UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

SISTEMA DE CLASIFICACIÓN AUTOMÁTICA DE CRÍTICAS DE CINE

PROYECTO FIN DE CARRERA

INGENIERÍA SUPERIOR DE TELECOMUNICACIÓN

Autora: Miriam Martín García

Tutor: Julio Villena Román

Julio 2009

Título: SISTEMA DE CLASIFICACIÓN AUTOMÁTICA DE CRÍTICAS DE CINE

Autor: Miriam Martín García

Tutor: Julio Villena Román

EL TRIBUNAL

Presidente:
Antonio de la Oliva Delgado

Secretario:
Damaris Fuentes Lorenzo

Vocal:
José Luis Martínez Fernández

Realizado el acto de defensa del Proyecto Fin de Carrera el día 10 de Julio de 2009 en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de:

Fdo: Presidente Fdo: Secretario Fdo: Vocal

AGRADECIMIENTOS

Quiero comenzar dando las gracias a mi tutor, Julio Villena Román, por su dedicación, ayuda y paciencia. Gracias por estar siempre dispuesto a resolver mis dudas y por haber hecho posible que esta etapa de mi vida llegue a su fin.

Quiero dar las gracias a toda mi familia, en especial a mis padres, mis hermanos y mis cuñados porque habéis sido testigos de todo el esfuerzo durante estos años y porque siempre me he sentido muy arropada por vosotros. Sé que tengo una familia genial a la que debo casi todo lo que soy y por eso me alegro mucho de poder compartir este momento con vosotros.

Quiero dedicárselo también a todos aquellos que ya no están a los que se que les habría hecho mucha ilusión verme terminar. En especial quiero dedicárselo a mi abuelo, al que siempre tengo muy presente y estoy segura que desde algún sitio está compartiendo, de alguna forma, este momento conmigo.

Gracias Dani por tu ayuda, tu constante apoyo y tu cariño que ha sido fundamental para terminar esta etapa, y sobre todo gracias por aportarme siempre el ánimo y la alegría para sacar ésta y muchas otras cosas adelante.

Gracias a todos los compañeros que han pasado por mi vida durante los años de universidad, por las horas compartidas y los grandes momentos vividos. En especial quiero agradecer a Virginia, Ana G., Ana R. y Natalia todas las risas y los buenos momentos compartidos desde el primer día hasta el último que han hecho que las horas de prácticas, de estudio y de esfuerzo fueran no solo llevaderas sino incluso divertidas, sin vosotras nada hubiera sido lo mismo.

Gracias a mis niñas de Miraflores (Cris, Tere, Ceci, Carmen y Miriam) por formar parte de mi vida desde siempre, por vuestro apoyo y vuestra amistad, por fin ha llegado el momento de celebrarlo chicas. Gracias también a mis compis de TID por haber vivido esta etapa final conmigo, por hacer del trabajo un sitio genial y porque en poco tiempo habéis logrado convertiros en grandes amigos.

Este proyecto se lo dedico especialmente a mi madre. Gracias por tu amor incondicional, por tu apoyo constante y tu comprensión. Creo que te lo mereces todo y por eso este proyecto quiero dedicártelo a ti.

RESUMEN

Considerada inicialmente una subdisciplina de la tarea de clasificación de documentos, en los últimos años la *clasificación de documentos basada en la opinión* (conocida en inglés bajo los nombres de *sentiment classification*, *sentiment analysis* u *opinion mining*) ha sido objeto de un creciente interés por parte de la comunidad de investigadores del procesamiento del lenguaje natural.

El creciente interés por el procesamiento automático de las opiniones contenidas en documentos de texto, es en parte consecuencia del aumento exponencial de contenidos generados por usuarios en la Web 2.0, y por el interés, entre otros, de empresas y administraciones públicas en analizar, filtrar o detectar automáticamente las opiniones vertidas por sus clientes o ciudadanos.

Este Proyecto de Fin de Carrera tiene como objetivo el diseño y la implementación de un sistema de clasificación automática de textos de opinión, concretamente de críticas cinematográficas vertidas por usuarios de internet, recogidas en diferentes webs dedicadas a tal fin. Los documentos serán clasificados, en una de las categorías definidas en el sistema (de acuerdo a la orientación afectiva de las críticas), aplicando diversas técnicas para el procesamiento del lenguaje natural (se aplicará en un caso el algoritmo kNN y en otro caso se hará uso de un diccionario afectivo). El hecho de conseguir un sistema automático de clasificación evitará la intervención humana y aumentará la rapidez con que se pueden procesar este tipo de documentos.

Con la realización de este proyecto, se comprobarán y analizarán también las dificultades encontradas en la implementación de un sistema de clasificación automática donde la naturaleza de los textos es de opinión.

ABSTRACT

As a subfield of document classification, *Opinion based document classification* (also known as sentiment classification, sentiment analysis or opinion mining) has been object of an increasing interest over the last years by the natural language research community.

This focus on automatic opinion detection in text documents is due to the exponential increase of contents produced by Web 2.0 users, as well as to the interest of companies and public administrations to be able to analyse, filter or detect opinions expressed by their clients or citizens.

The aim of this project is the design and implementation of an automatic opinion classification system, specifically, the classification of film reviews written by internet users that have been collected among different specialized websites. The documents will be classified into one of the defined system's categories (according to the review's affective orientation), applying diverse techniques for the natural language processing (both a kNN algorithm and an affective dictionary will be used). Such a kind of automatic classification system avoids any human intervention and considerably decreases the document's manipulation time.

Problems and difficulties found while implementing the system will be thoroughly commented and analysed.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. MOTIVACIÓN

La clasificación es un concepto bien conocido por quienes se dedican a la documentación. Sin entrar en disquisiciones formales, se trata de organizar los documentos en alguna forma que permita después su mejor recuperación. En torno a ello se han elaborado diversas técnicas, que se han aplicado con mejor o peor fortuna en unas y otras aplicaciones. Con la creciente disponibilidad de documentos en formato electrónico, susceptibles, por consiguiente, de ser procesados de manera automática, surge la posibilidad de abordar la clasificación de documentos de manera automática.

El problema de clasificar automáticamente, según el tema del que se trate, el creciente volumen de información disponible es un área de investigación ampliamente estudiada durante años. Sin embargo, clasificar automáticamente textos de acuerdo al sentimiento que contienen, y no de acuerdo a su temática, es un tema muy de actualidad en cuya investigación se están haciendo numerosos avances.

Considerada inicialmente una subdisciplina de la tarea de clasificación de documentos, en los últimos años la *clasificación de documentos basada en la opinión* (conocida en inglés bajo los nombres de *sentiment classification*, *sentiment analysis* u *opinion mining*) ha sido objeto de un creciente interés por parte de la comunidad de investigadores del procesamiento del lenguaje natural. Si en la tarea de clasificación de documentos clásica el problema consiste en decidir la temática de un documento de entre un conjunto de temáticas posibles (por ejemplo, centrándose en el ámbito de las noticias periodísticas, distinguir cuando nos encontramos ante un texto de política, sociedad o deportes), en la clasificación basada en la opinión se trata, por ejemplo, de determinar si en un texto se expresan opiniones negativas o positivas acerca del tema principal del que trata dicho texto.

El creciente interés por el procesamiento automático de las opiniones contenidas en documentos de texto, es en parte consecuencia del aumento exponencial de contenidos generados por usuarios en la Web 2.0, y por el interés, entre otros, de empresas y administraciones públicas en analizar, filtrar o detectar automáticamente las opiniones vertidas por sus clientes o ciudadanos.

La gran cantidad de opiniones que los usuarios emiten sobre las características de los productos en blogs, foros y en documentos en Internet, son de gran ayuda para los posibles compradores o para las compañías que los producen. Sin embargo, determinar de forma automática si un usuario tiene una opinión positiva o negativa de las características de un producto o del propio producto es un problema complejo que requiere de varios pasos para su resolución.

Como ya se ha comentado, los sistemas de tratamiento de información que realizan minería de opiniones, son un tema actual y muy activo de investigación y desarrollo que tienen una variedad de aplicaciones, que van desde el análisis automatizado de opiniones de películas, obras, o productos en general, hasta estudios que permitan dar seguimiento a cómo va evolucionando la percepción de los ciudadanos acerca de las empresas, gobernantes o políticos. Es importante mencionar que debido a todas las posibles aplicaciones, hay un buen número de empresas, tanto grandes como PYMES, que tiene la minería de opiniones y el análisis de emociones como una de sus misiones.

1.2. OBJETIVOS

El objetivo de este Proyecto de Fin de Carrera es el diseño y la implementación de un sistema de clasificación automática de textos de opinión, concretamente de críticas cinematográficas vertidas por usuarios de internet, recogidas en diferentes webs dedicadas a tal fin. Los documentos serán clasificados, en una de las categorías definidas en el sistema, aplicando diversas técnicas para el procesamiento del lenguaje natural. El hecho de conseguir un sistema automático de clasificación evitará la intervención humana y aumentará la rapidez con que se pueden procesar este tipo de documentos.

Tomando como punto de partida trabajos de otros autores para el inglés, en este proyecto se expondrán los resultados obtenidos en la experimentación con un clasificador supervisado de documentos basado en la opinión para el español. Como paso previo a la experimentación, y ante la ausencia de recursos adecuados, a nuestro entender, en español para desarrollar el trabajo, se presentará un corpus propio elaborado de críticas de cine en español.

La decisión de implementar un clasificador empleando como documentos de trabajo **críticas de cine** se antoja muy conveniente para analizar un sistema capaz de detectar el sentimiento o la polaridad en textos de opinión; existen muchas colecciones de este tipo de críticas que pueden encontrarse con facilidad en la red y además la variedad de opiniones, incluso acerca de una misma película, es muy elevada lo cual facilita enormemente la tarea de encontrar muchos textos para las distintas categorías con las que se desea trabajar.

Con la realización de este proyecto, se comprobarán y analizarán también las dificultades encontradas en la implementación de un sistema de clasificación automática donde la naturaleza de los textos es de opinión y estos han sido elaborados por usuarios comunes de Internet, los cuales pueden expresan libremente cualquier sentimiento en un lenguaje coloquial.

Más concretamente, los principales objetivos buscados con el desarrollo de este proyecto se enumeran a continuación:

- Construir un corpus de documentos adecuado para la implementación de un sistema de clasificación automática de textos de opinión como el que se quiere desarrollar.
- Ser capaces de extraer la información relevante contenida en textos de opinión para obtener una representación estructurada de los mismos que facilite su procesado y análisis.
- Lograr que el conocimiento adquirido a partir de unos documentos ya categorizados permita desarrollar un sistema de clasificación automática válido y eficiente ante una consulta del usuario.

- Entender y analizar los problemas que se puedan derivar del empleo de textos de opinión, o textos con una carga afectiva significativa, a la hora de implementar un clasificador automático de documentos.

1.3. ESTRUCTURA DEL DOCUMENTO

El presente documento se estructura en diversos capítulos cuyos contenidos se detallan brevemente a continuación:

- **Capítulo 1**: *Introducción*

Tras una breve explicación acerca de los fundamentos y las motivaciones sobre las que se apoya un sistema de clasificación automática de textos (y más concretamente un sistema de clasificación de textos de opinión), se enumeran los objetivos perseguidos con el desarrollo de este Proyecto de Fin de Carrera y se describe la estructura de la memoria realizada.

- **Capítulo 2:** *Estado del arte*

En este capítulo se realiza un repaso acerca de algunos de los conceptos previos que pueden resultar de importancia a la hora de abordar el diseño de un clasificador. Por un lado se detallan algunos conceptos acerca de la clasificación automática de documentos clásica (centrada en el tema) y por otro lado se explica el estado del arte en cuanto a la clasificación de documentos de opinión se refiere.

- **Capítulo 3:** *Arquitectura del sistema*

Se realiza una descripción general sobre la arquitectura empleada para el tipo de clasificador de documentos que se desea implementar. Se detallan, grosso modo, las fases generalmente necesarias para el diseño de un clasificador automático de textos, comentando algunas de las técnicas que pueden emplearse en cada una de las fases.

- **Capítulo 4**: *Diseño e implementación del sistema*

Se describen con detalle todas las decisiones de diseño tomadas así como los pasos seguidos para la implementación de cada uno de los dos sistemas desarrollados en este proyecto; el basado en el algoritmo de los k vecinos más próximos y el basado en el empleo de un diccionario afectivo.

- **Capítulo 5**: *Validación del sistema*

En este capítulo se detalla el corpus empleado para la implementación de los dos sistemas y las medidas necesarias para su evaluación. Se presentan las pruebas realizadas, en ambos casos, y los resultados obtenidos tras su ejecución.

- **Capítulo 6**: *Conclusiones y trabajos futuros*

En este último capítulo, se presentan las conclusiones alcanzadas tras la realización del proyecto y el análisis de los resultados obtenidos. Por otra parte, se presentan posibles líneas de investigación en las que seguir trabajando con el fin de obtener mejoras en el sistema y ahondar más en el ámbito de la clasificación automática de textos.

2. ESTADO DEL ARTE

2.1. CLASIFICACIÓN AUTOMÁTICA DE TEXTOS

2.1.1. INTRODUCCIÓN

Durante los últimos veinte años, la rápida expansión que ha experimentado Internet en todo el mundo ha hecho posible que el acceso a todo tipo de información sea una tarea de baja complejidad. Cada vez es mayor el número de fuentes de contenidos y el volumen de datos que se tiene al alcance y este crecimiento explosivo de documentos disponibles complica su exploración y análisis. Por consiguiente, son necesarios nuevos métodos que ayuden a los usuarios a filtrar y estructurar la información relevante. Por ello, poder organizar la información de forma automática ha pasado a ser una tarea de vital importancia y llevar a cabo una gestión eficiente de la información se ha convertido en algo imprescindible. Por este motivo cada vez son más necesarias herramientas que puedan automatizar esta clasificación.

La clasificación automática de texto consiste en un conjunto de algoritmos, técnicas y sistemas capaces de asignar un documento a una o varias categorías o grupos de documentos, construidos según su afinidad temática. Para ello (**Figura 1**) se emplean técnicas de Aprendizaje Automático (*ML, Machine Learning*) y de Procesamiento de Lenguaje Natural (*NLP, Natural Language Processing*).

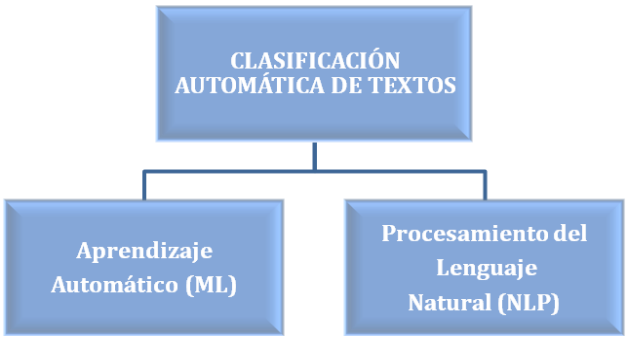

Figura 1. Clasificación automática de texto

El Procesamiento de Lenguaje Natural (PLN) estudia los problemas inherentes al procesamiento y manipulación de lenguajes naturales, haciendo uso de ordenadores. Pretende adquirir conocimiento sobre el modo en que los humanos entienden y utilizan el lenguaje, de tal forma que se pueda llevar a cabo el desarrollo de herramientas y técnicas para conseguir que los ordenadores puedan entenderlo y manipularlo. Sus fundamentos residen en un conjunto muy amplio de disciplinas: ciencias de la información y los computadores, lingüística, matemáticas, ingeniería eléctrica y electrónica, inteligencia artificial y robótica, psicología, etc. Existe un gran número de aplicaciones donde el PLN resulta de gran utilidad (traducción máquina, procesamiento y resumen de textos escritos en lenguaje natural, interfaces de usuario, reconocimiento de voz, etc.).

Para el diseño de la función de clasificación se pueden emplear diferentes técnicas de aprendizaje, debiendo disponer para ello de un conjunto de documentos (conjunto de entrenamiento), que previamente han sido clasificados dentro de una determinada categoría. Estos algoritmos de aprendizaje o entrenamiento requieren una representación estructurada de los documentos. La más empleada es la basada en el modelo de espacio vectorial, donde cada documento se transforma en un vector de palabras clave a las que se les asigna un peso en función de la importancia o relevancia que éstas representen dentro del documento. Una vez que el clasificador ha sido entrenado con el correspondiente grupo de textos, su efectividad se evalúa comparando las categorías que ha asignado a los documentos del set de prueba con las que éstos ya tenían asignadas. Este esquema permite alcanzar una precisión comparable a la obtenida por expertos humanos, reduciendo así los costes de mano de obra.

Algunos ejemplos de los entornos en los que se emplea la clasificación automática son: indexación automática de textos, filtrado de textos, clasificación de páginas Web, filtrado de correos electrónicos (*spam*), o clasificación de noticias.

2.1.2. TIPOS DE CLASIFICADORES

En el esquema de la **Figura 2** se presentan las distintas características que puede presentar un clasificador, según diferentes puntos de vista:

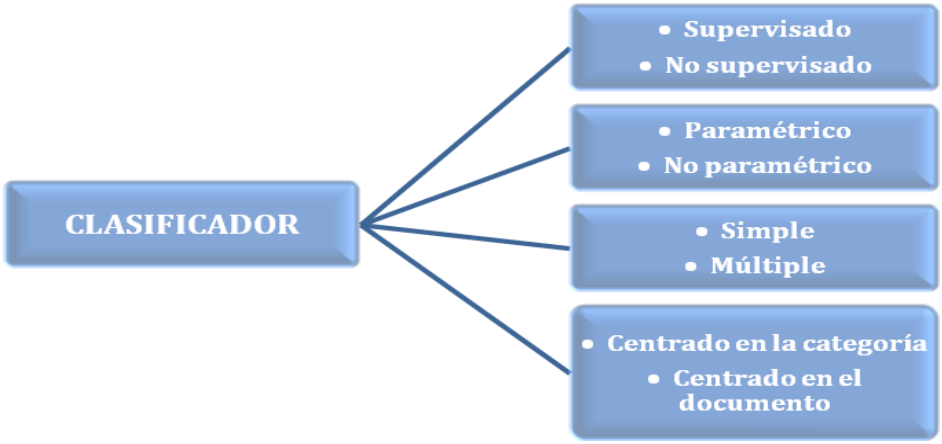

Figura 2. Tipos de clasificadores

2.1.2.1 Clasificación supervisada y no supervisada

Clasificación supervisada: partiendo de una serie de categorías conceptuales prediseñadas a priori, se encarga de asignar cada documento a la categoría correspondiente. Requiere la elaboración manual o intelectual del conjunto de categorías. Además, es necesaria una fase de entrenamiento por parte del clasificador.

El objetivo que se persigue en los clasificadores supervisados es el siguiente: elaborar un patrón representativo para cada una de las categorías entrenadas y aplicar alguna función que permita estimar la similitud entre el documento a clasificar y cada uno de estos patrones. Aquel patrón o patrones que presenten más concordancias con el documento indicarán la categoría o categorías a las que pertenece el mismo. El proceso de elaboración de los patrones necesita un conjunto de documentos previamente clasificados y se conoce como aprendizaje o entrenamiento.

Clasificación no supervisada: no existen categorías previas o cuadros de clasificación establecidos a priori. Los documentos se clasifican en función de su contenido, de forma automática, sin asistencia manual. Es una segmentación o agrupamiento automático, en inglés conocido como *clustering*.

2.1.2.2 Clasificación paramétrica y no paramétrica

Clasificación paramétrica: en el entrenamiento de un clasificador paramétrico se emplea el *set* de entrenamiento para estimar o aprender los parámetros del modelo. El *set* de test que contiene documentos a clasificar se emplea para determinar la capacidad de generalización del clasificador [Tumer & Ghosh, 1995].

Clasificación no paramétrica: se subdivide en dos categorías. La primera está basada en patrones y se obtiene una descripción de cada categoría en términos de un patrón, normalmente en forma de vector de términos con peso, como por ejemplo el clasificador Rocchio. La clasificación de los documentos se realiza en función de las similitudes existentes entre cada documento y los distintos patrones [Bacan, Pandzic, & Gulija, 2005]. La segunda categoría está basada en ejemplos y los documentos se clasifican según las similitudes que presenten con ejemplos del conjunto de entrenamiento. El clasificador más conocido es el del vecino más cercano (*KNN, K-Nearest Neighbour*).

2.1.2.3 Clasificación simple y clasificación múltiple

Clasificación simple: cada documento tiene una única categoría. Se trata de una clasificación donde las categorías no se solapan. Un caso especial es la clasificación binaria, donde cada documento pertenece a una categoría o a su complementaria.

Clasificación múltiple: cada documento puede recibir un número variable de categorías. En este caso, las categorías sí se pueden solapar.

2.1.2.4 Clasificación centrada en la categoría y en el documento

Una vez construido el clasificador existen dos formas en las que puede ser utilizado, teniendo en cuenta el hecho de que el conjunto de categorías C o el conjunto de documentos D puede que no se encuentren disponibles de forma completa desde el comienzo [Sebastiani, 2002].

Clasificación Centrada en la Categoría (CPC, Category-Pivoted Classification)**:** dado el documento se pueden encontrar todas las categorías dentro de las cuales se puede clasificar.

Clasificación Centrada en el Documento (DPC, Document-Pivoted Classification): dada la categoría debemos encontrar todos los documentos que pueden ser clasificados dentro de ella. Adecuada cuando una nueva categoría se añade al conjunto después de que varios documentos ya hayan sido clasificados y es necesario que dichos documentos se vuelvan a tener en cuenta para una posible clasificación.

2.1.3. TÉCNICAS DE CLASIFICACIÓN AUTOMÁTICA DE TEXTOS

2.1.3.1 Algoritmos probabilísticos

Se basan en la teoría probabilística, en especial en el teorema de Bayes, el cual permite estimar la probabilidad de un suceso a partir de la probabilidad de que ocurra otro suceso, del cual depende el primero. El algoritmo más conocido, y también el más simple, es el denominado *Naïve Bayes* [Figuerola, Alonso Berrocal, Zazo Rodríguez, & Rodriguez, 2004], que estima la probabilidad de que un documento pertenezca a una categoría. Dicha pertenencia depende de la posesión de una serie de características, de cada una de las cuales se conoce la probabilidad de que aparezcan en los documentos que pertenecen a la categoría en cuestión. Naturalmente, dichas características son los términos que conforman los documentos, y tanto su probabilidad de aparición en general, como la probabilidad de que aparezcan en los documentos de una determinada categoría, pueden obtenerse a partir de los documentos de entrenamiento; para ello se utilizan las frecuencias de aparición en la colección de entrenamiento.

Cuando las colecciones de aprendizaje son pequeñas, pueden producirse errores al estimar dichas probabilidades. Por ejemplo, cuando un determinado término no aparece nunca en esa colección de aprendizaje pero aparece en los documentos a categorizar. Esto implica la necesidad de aplicar técnicas de suavizado, a fin de evitar distorsiones en la obtención de las probabilidades [Figuerola, Alonso Berrocal, Zazo Rodríguez, & Rodriguez, 2004].

Con dichas probabilidades, obtenidas de la colección de entrenamiento, podemos estimar la probabilidad de que un nuevo documento, dado que contiene un conjunto determinado de términos, pertenezca a cada una de las categorías. La más probable, obviamente, es a la que será asignado.

2.1.3.2 Algoritmo de Rocchio

El llamado *algoritmo de Rocchio* [Figuerola, Alonso Berrocal, Zazo Rodríguez, & Rodriguez, 2004] se aplica en la realimentación de consultas. Una vez formulada y ejecutada una primera consulta, el usuario examina los documentos que el clasificador ha devuelto y determina cuáles le resultan relevantes y cuáles no. Con estos datos, el sistema genera automáticamente una nueva consulta, basándose en los documentos que el usuario señaló como relevantes o no relevantes. En este contexto, el algoritmo de Rocchio proporciona un sistema para construir el vector de la nueva consulta, recalculando los pesos de los términos de ésta y aplicando un coeficiente a los pesos de los la consulta inicial, otro a los de los documentos relevantes y otro distinto a los de los no relevantes.

En el ámbito de la categorización, el mismo algoritmo de Rocchio proporciona un sistema para construir los patrones de cada una de las clases o categorías de documentos. Así, partiendo de una colección de entrenamiento, previamente categorizada de forma manual, y aplicando el modelo vectorial, se pueden construir vectores patrón para cada una de las categorías, considerando como ejemplos positivos los documentos de entrenamiento de esa categoría, y como ejemplos negativos los de las demás categorías.

Una vez que se tienen los patrones de cada una de las clases, el proceso de entrenamiento o aprendizaje está completado. Para categorizar nuevos documentos, simplemente se estima la similitud entre el nuevo documento y cada uno de los patrones. El que presenta un índice mayor indica la categoría a la que se debe asignar ese documento (**Figura 3**).

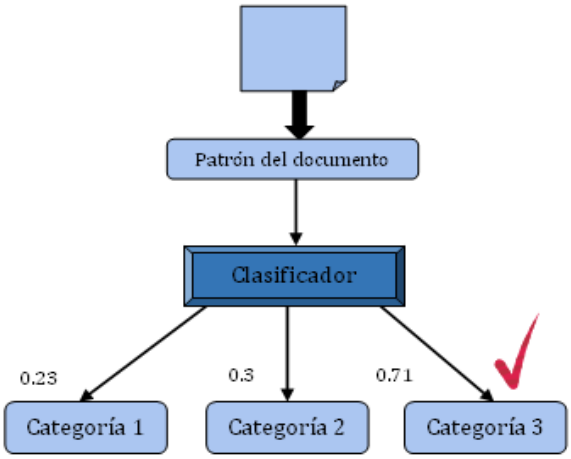

Figura 3. Clasificador Rocchio

2.1.3.3 Algoritmos del vecino más próximo y variantes

El algoritmo del vecino más próximo (*Nearest Neighbour, NN*) es uno de los más sencillos de implementar. La idea básica es como sigue: si se calcula la similitud entre el documento a clasificar y cada uno de los documentos de entrenamiento, aquél de éstos más parecido estará indicando a qué clase o categoría se debe asignar el documento que se desea clasificar.

Una de las variantes más conocidas de este algoritmo es la del ***k-nearest neighbour*** o ***kNN*** que consiste en tomar los *k* documentos más parecidos, en lugar de sólo el primero. Como en esos *k* documentos los habrá, presumiblemente, de varias categorías, se suman los coeficientes de los de cada una de ellas. La que más puntos acumule, será la candidata idónea. El *kNN* une a su sencillez una eficacia notable. Obsérvese que el proceso de entrenamiento no es más que la indexación o descripción automática de los documentos, y que tanto dicho entrenamiento como la propia categorización pueden llevarse a cabo con instrumentos bien conocidos y disponibles para cualquiera. De otra parte, numerosas pruebas experimentales han mostrado su

eficacia. *KNN* parece especialmente eficaz cuando el número de categorías posibles es alto, y cuando los documentos son heterogéneos y difusos.

Es un método de clasificación no paramétrico, ya que no se hace ninguna suposición distribucional acerca de las variables predictoras. Para inferir la categoría de un ejemplo desconocido, el algoritmo compara ese ejemplo con todos los ejemplos de entrenamiento, calculando la distancia entre ellos. A continuación, la clase mayoritaria de entre los *k* ejemplos más similares al de entrada es la categoría inferida para el mismo. Generalmente se usa la distancia Euclídea:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{|C|}(x_{ik} - x_{jk})^2}$$ **Ecuación 1**

En la **Figura 4** se muestra un ejemplo del algoritmo kNN para un sistema de dos atributos. En este ejemplo se ve cómo en el proceso de aprendizaje se almacenan todos los ejemplos de entrenamiento. Se han representado los ejemplos de acuerdo a los valores de sus dos atributos y la clase a la que pertenecen (las clases son *positivo* y *negativo*). La clasificación consiste en la búsqueda de los *k* ejemplos (en este caso tres) más cercanos al ejemplo a clasificar. Concretamente, el ejemplo *a* se clasificaría como *negativo*, y el ejemplo *b* como *positivo*.

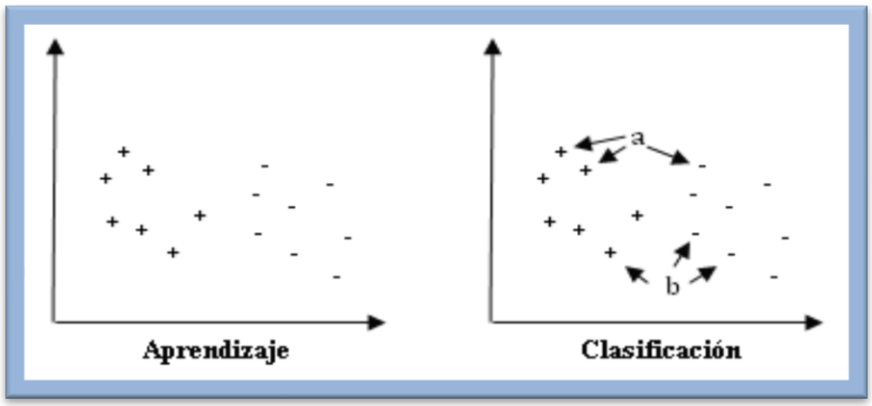

Figura 4. Algoritmo KNN

2.1.3.4 Árboles de clasificación

Es uno de los métodos de aprendizaje inductivo supervisado no paramétrico más utilizado [Cortijo Bon, 2000]. Como forma de representación del conocimiento, los árboles de clasificación destacan por su sencillez. A pesar de que carecen de la expresividad de las redes semánticas o de la lógica de primer orden, su dominio de aplicación no está restringido a un ámbito concreto sino que pueden utilizarse en diversas áreas: diagnóstico médico, juegos, predicción meteorológica, control de calidad, etc.

Un árbol de clasificación es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir del conjunto de prototipos (documentos). Esta partición recursiva se traduce en una *organización jerárquica* del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada *nodo interior* contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada *nodo hoja* se refiere a una decisión (clasificación).

La clasificación de patrones se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezado por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja. La etiqueta asignada a esta hoja es la que se asignará al patrón a clasificar.

La metodología a seguir puede resumirse en dos pasos y se esquematiza en la **Figura 5**:

Aprendizaje: Consiste en la construcción del árbol a partir de un conjunto de prototipos. Constituye la fase más compleja y la que determina el resultado final. A esta fase se dedica la mayor parte de la atención.

Clasificación: Consiste en el etiquetado de un patrón, *X*, independiente del conjunto de aprendizaje. Se trata de responder a las preguntas asociadas a los nodos interiores utilizando los valores de los atributos del patrón *X*. Este proceso se repite desde el nodo raíz hasta alcanzar una hoja, siguiendo el camino impuesto por el resultado de cada evaluación.

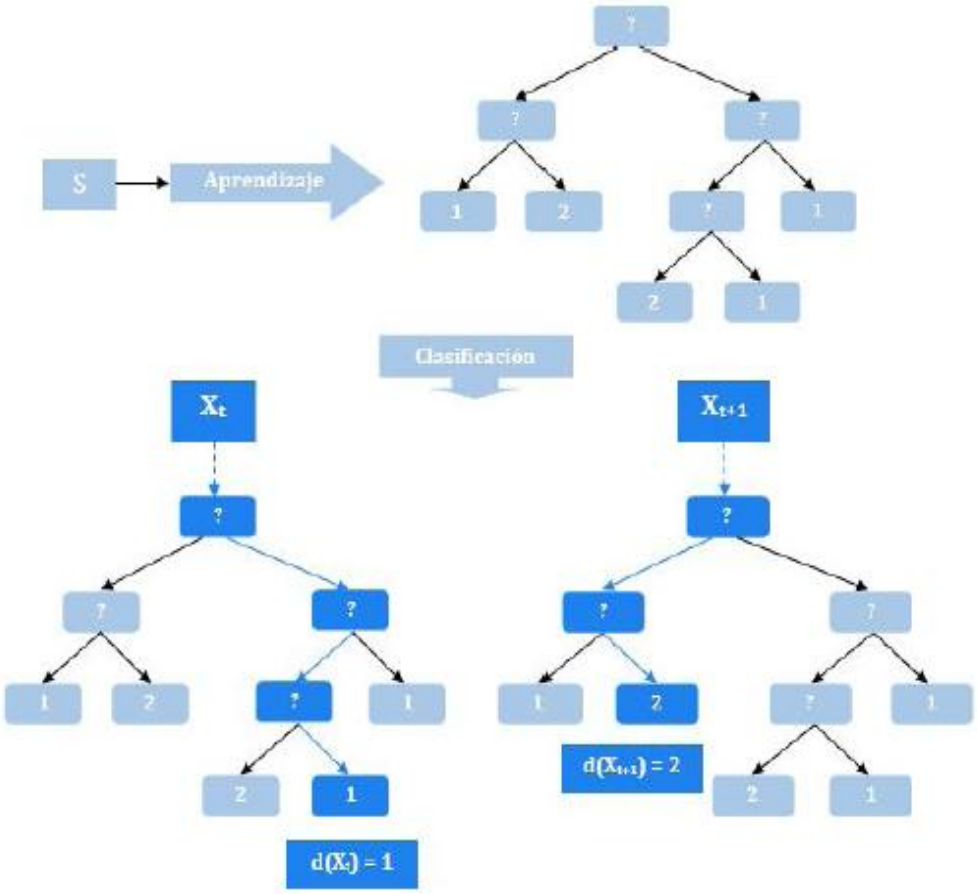

Figura 5. Árboles de clasificación

2.1.3.5 Máquinas de Vectores Soporte

Las Máquinas de Vectores Soporte (o SVM de 'Support Vector Machine') son procedimientos de clasificación y regresión basados en la teoría estadística del aprendizaje [Vapnik, 1995]. Se puede definir una SVM como una clase específica de algoritmos preparados para el entrenamiento eficaz de una máquina de aprendizaje lineal en un espacio inducido por una función núcleo (o kernel), de acuerdo a unas reglas de generalización empleando técnicas de optimización.

Las dos ideas fundamentales para la construcción de un clasificador SVM son la transformación del espacio de entrada en un espacio de alta dimensión y la localización en dicho espacio de un hiperplano separador óptimo. La transformación inicial se realiza mediante la elección de una función kernel adecuada. La ventaja de trabajar en un

espacio de alta dimensión radica en que las clases consideradas serán linealmente separables con alta probabilidad, y por tanto, encontrar un hiperplano separador óptimo será poco costoso desde el punto de vista computacional. Además, dicho hiperplano vendrá determinado por unas pocas observaciones, denominadas, vectores soporte por ser las únicas de las que depende la forma del hiperplano. Una de las principales dificultades en la aplicación de este método radica en la elección adecuada de la función kernel. Es decir, construir la función de transformación del espacio original a un espacio de alta dimensión es un punto crucial para el buen funcionamiento del clasificador.

La forma final de la regla de clasificación para un clasificador binario (dos clases +1 y -1) queda como sigue:

$$f(x) = b + \sum_i \propto_i K(x, x_i)$$ **Ecuación 2**

donde b y α_i son parámetros aprendidos por el clasificador durante el proceso de entrenamiento, $K(x, x_i)$ es el valor de la función kernel para los puntos x y x_i . Si f(x) es mayor que un umbral entonces la emoción estimada para un punto x será una +1 y será -1 en caso contrario.

2.1.3.6 Redes neuronales

Las redes neuronales han sido propuestas en numerosas ocasiones como instrumentos útiles para la clasificación automática. De una manera genérica, una de las principales aplicaciones de las redes neuronales es el reconocimiento de patrones. Por tanto, no es de extrañar que se hayan aplicado a problemas de categorización de documentos [Figuerola, Alonso Berrocal, Zazo Rodríguez, & Rodriguez, 2004].

Una red neuronal (**Figura 6**) consta de varias capas de unidades de procesamiento o neuronas interconectadas. En el ámbito que nos ocupa, la capa de entrada recibe términos, mientras que las unidades o neuronas de la capa de salida mapean clases o categorías.

Una neurona es un dispositivo sencillo formado por una serie de entradas y una única salida [Cabello Pardos, 2004]. Cada neurona acepta como entrada las salidas procedentes de otras neuronas, siendo la entrada efectiva a la neurona la suma ponderada de las entradas reales a dicha neurona. Cada neurona se caracteriza por su estado de activación, que es un valor que oscila entre 0 (no activada) y 1. Cada neurona realiza una tarea sencilla: recibe la información de entrada de las neuronas o del exterior y la usa para calcular una señal de salida que se propaga a otras unidades.

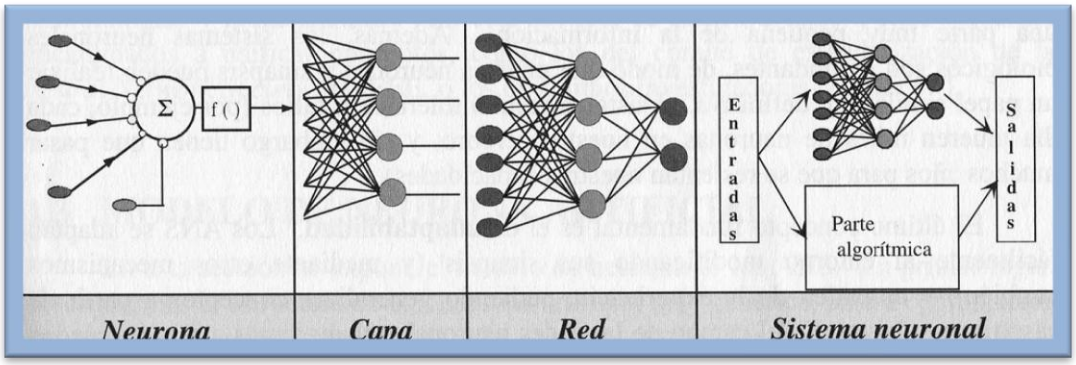

Figura 6. Descomposición de una red neuronal

Las interconexiones tienen pesos, es decir, un coeficiente que expresa la mayor o menor fuerza de la conexión. Es posible entrenar una red para que, dada una entrada determinada (los términos de un documento), produzca la salida deseada (la clase que corresponde a ese documento). El proceso de entrenamiento consta de un ajuste de los pesos de las interconexiones, a fin de que la salida sea la deseada.

2.2. ANÁLISIS DE OPINIÓN O CLASIFICACIÓN AFECTIVA

2.2.1. INTRODUCCIÓN

Considerada inicialmente una subdisciplina de la tarea de clasificación de documentos, en los últimos años la clasificación de documentos basada en la opinión (conocida en inglés bajo los nombres de *sentiment classification*, *sentiment analysis* u *opinion mining*) ha sido objeto de un creciente interés por parte de la comunidad de investigadores del PLN.

Si en la tarea de clasificación de documentos clásica el problema consiste en decidir la temática de un documento de entre un conjunto de temáticas posibles, en la clasificación basada en la opinión se trata de determinar, por ejemplo, si en los textos se expresan opiniones negativas o positivas. Es desde este prisma, considerando "opinión negativa" y "opinión positiva" como las dos clases de salida de la tarea, bajo el que se considera que la clasificación basada en la opinión es una subdisciplina de la clasificación de documentos. Sin embargo, la naturaleza subjetiva de los documentos con los que se trabaja (análisis de productos, críticas de cine o música, intervenciones políticas, contenidos generados por internautas como blogs o foros, etc.) añaden dificultad a la tarea y hacen necesario plantear soluciones distintas a las utilizadas en la clasificación de documentos clásica. En la clasificación basada en la opinión entran en juego fenómenos del lenguaje no sólo léxicos, sintácticos y semánticos, sino pragmáticos y en gran medida de conocimiento del mundo, es decir, que las técnicas de PLN tienen mucho más peso y juegan un papel más importante.

La clasificación orientada a la opinión puede variar desde la categorización de la polaridad del sentimiento en críticas, hasta la determinación de la intensidad de las opiniones en artículos de prensa para identificar las perspectivas en debates políticos, o el análisis del humor en blogs. La parte que es verdaderamente interesante en torno a estos problemas son los nuevos retos y oportunidades que con ellos se ofrecen.

2.2.2. APLICACIONES

Como ejemplos específicos se consideran, por ejemplo, las siguientes aplicaciones del análisis automático de emociones:

2.2.2.1 Búsqueda en foros y blogs

Los foros online y los blogs son un medio de comunicación cada vez más extendido para el intercambio de experiencias, opiniones y reclamaciones o quejas. La búsqueda de palabras clave en foros puede ayudar a encontrar mensajes generalmente relacionados con áreas de interés subjetivas. Puede ser muy útil ser capaz de buscar o filtrar mensajes basados en sus connotaciones positivas o negativas con respecto un tema en particular.

2.2.2.2 Política

Como es bien sabido, las opiniones son de gran importancia en política. Algunos trabajos se han centrado en entender qué están pensando los votantes, mientras que otros proyectos tienen como objetivo fundamental aclarar las posturas de los políticos, como a qué figuras públicas apoyan o se oponen, para mejorar la calidad de la información a la que los votantes tienen acceso. El análisis de emociones ha sido propuesto específicamente como una tecnología clave para posibilitar el llamado *eRuleMaking*, permitiendo el análisis automático de opiniones que la gente sostiene sobre políticas pendientes o sobre propuestas de regulaciones gubernamentales.

2.2.2.3 Clasificación de emails y distribución priorizada

En un servicio de información y reclamaciones dado, los mensajes electrónicos son típicamente clasificados por tema con el objetivo de derivar los mensajes apropiadamente hacia el agente más cualificado para ellos. Puede ser de gran valor para las compañías mejorar el sistema de asignación de rutas de mensajes basado en temas, detectando la presencia de sentimientos negativos (enfado, frustración) en el mensaje del cliente. Los beneficios potenciales incluyen mejorar la retención de clientes y la recopilación de información sobre la naturaleza de contactos previos con el servicio al cliente.

2.2.2.4 Inteligencia de negocio o inteligencia competitiva

El campo de *opinion mining* y *sentiment analysis* es el idóneo para muchos tipos de aplicaciones de inteligencia. Es más, la inteligencia de negocio parece ser uno de los factores principales para el interés empresarial.

La clasificación de foros, blogs y datos contenidos en emails por su actitud o emoción puede servir a los propósitos de la inteligencia de negocio y de mercado. ¿Cuál es la opinión en conjunto de la calle con respecto a un determinado producto, servicio o compañía? ¿Cuáles son las citas claves o frases que realmente ilustran estas opiniones? ¿Cómo se propagan estas opiniones por la red y que influencia tienen?

Se considera, por ejemplo, el siguiente escenario [Boo Pang y Lillian Lee, 2008]: Un importante fabricante de ordenadores está decepcionado con la inesperada bajada de las ventas, y se encuentra preguntándose a sí mismo: "*¿Por qué los consumidores no están comprando nuestros ordenadores portátiles*?". Aunque los datos concretos como el peso o el precio de los portátiles de un competidor son muy relevantes, la respuesta a esta pregunta requiere centrarse más en el punto de vista personal de la gente que en características objetivas. Además las opiniones subjetivas que se refieren a cualidades intangibles (por ejemplo "*el diseño es hortera*" o "*el servicio al consumidor era condescendiente*") o incluso a malentendidos (por ejemplo "*las actualizaciones de los drivers no están disponibles*" cuando dicho dispositivo de drivers sí existe en realidad) deben tenerse en cuenta también.

Las tecnologías de análisis de emociones para la extracción de opiniones procedentes de documentos no estructurados pueden ser excelentes herramientas para el manejo de tareas relacionadas con la que se acaba de describir. Continuando con el escenario de ejemplo: sería difícil tratar de analizar directamente a los compradores de portátiles que no han comprado el producto de la compañía. Preferiblemente, se puede emplear un sistema que (a) encuentre críticas u otras expresiones de opinión en la Web (grupos de noticias, blogs y webs de opiniones son probables fuentes productivas) y entonces (b) crear versiones agrupadas de críticas individuales. Esto puede salvar a un analista de tener que leerse potencialmente docenas o incluso centenares de versiones de las mismas quejas.

Nótese que las fuentes de Internet pueden variar sustancialmente en forma, contenido e incluso gramaticalmente; este hecho enfatiza la necesidad de técnicas robustas aun cuando sólo se considere un único idioma.

2.2.2.5 Aplicaciones como un subcomponente tecnológico

Los sistemas de análisis afectivo y minería de opinión también tienen un papel potencialmente importante capacitando tecnologías para otros sistemas.

Una posibilidad es aplicarlo como una mejora de los sistemas de recomendación ya que se comportan como un sistema que no recomendará ítems que reciban mucho feedback negativo.

En sistemas online que muestren anuncios, es una ayuda para detectar páginas Web que contengan un contenido sensiblemente inapropiado para ser colocado; para sistemas más sofisticados, puede ser útil para sacar anuncios de productos cuando se detecten sentimientos relevantemente positivos, y quizás más importante, para vetar el anuncio cuando se detecten declaraciones claramente negativas.

En general, el tratamiento computacional de las emociones ha sido motivado en parte por el deseo de mejorar la interacción entre los humanos y los ordenadores.

Con este tipo de aplicaciones en mente, existen multitud de líneas de investigación específicas en el análisis de emociones. El análisis de opinión, en concreto la distinción entre declaraciones objetivas y subjetivas, opera a nivel de oración o frase. La clasificación afectiva a nivel de documentos busca determinar si un texto indica una opinión positiva o muestra apoyo o, por el contrario, indica una opinión negativa o se muestra en contra en relación al tema de discusión. En particular, las críticas de películas y productos han recibido una gran atención en lo que concierne a esta tarea.

2.2.3. MODELOS EMOCIONALES Y TÉCNICAS EMPLEADAS

A continuación se presenta un breve resumen sobre la investigación en el ámbito de la identificación emocional de textos. Primero, se describen los modelos psicológicos emocionales que mejor se adaptan al mundo computacional. Seguidamente, se detallan las técnicas más utilizadas para la detección automática de emociones a partir de texto.

2.2.3.1 Modelos emocionales

A continuación se revisan los modelos psicológico-computacionales más relevantes aplicables al ámbito del análisis de opinión.

El modelo más intuitivo para representar emociones es el **basado en categorías emocionales**, como pueden ser alegría, tristeza, ira, etc. [Plutchik y Kellerman, 1980; Ekman, 1993]. Una mejora de éste, es el **modelo Circumplex** [Schlosberg, 1952], que utiliza una circunferencia con dos ejes que representan sendas características emocionales, dando lugar a diferentes versiones según cuales sean: valencia (positivo/negativo) o activación y control (excitado/tranquilo) [Rusell, 1980].

Asimismo, existe un modelo similar al Circumplex llamado **dimensiones emocionales** [Schlosberg, 1954], que cuantifica las dimensiones de valencia, activación y control (dominado/dominante) mediante un vector de tres elementos. Finalmente, y en contraposición a los modelos anteriores, existe **el modelo OCC** [Ortony, Clore, y Collins, 1988], que presenta una jerarquía cognitiva de las emociones evitando el uso de categorías y dimensiones.

2.2.3.2 Algunas técnicas empleadas para el análisis afectivo de textos

A continuación se presentan algunas de las técnicas empleadas en el análisis de documentos de opinión las cuales tienen como objetivo, basándose en diferentes enfoques, la detección y clasificación de emociones presentes en los textos.

- **Basadas en la recuperación de información**

Una de las técnicas empleadas para determinar la orientación semántica de un texto es hacer un análisis PMI-IR (Pointwise Mutual Information and Information Retrieval).

En [Turney, 2002] se describe un clasificador no supervisado basado en la opinión que decide el carácter positivo o negativo de un documento en base a la orientación semántica de los términos que aparecen en el mismo. Esta orientación se calcula mediante el algoritmo PMI-IR comentado, que consiste en estimar la Información Mutua Puntual (*Pointwise Mutual Information*) entre el término en cuestión y un par de palabras "semilla" que sirven de representantes inequívocos. De manera intuitiva, la idea detrás de este cálculo de la orientación semántica es que expresiones que indiquen una opinión positiva aparecerán con mayor frecuencia cerca de una palabra con claras connotaciones positivas como *excelente* y con mucha menor frecuencia cerca de una palabra con connotaciones negativas como *horrible*. Dicho de otra forma, palabras que expresen un sentimiento parecido a menudo co-aparecen en un mismo texto, mientras que palabras que expresan un sentimiento contrapuesto raramente aparecen juntas. De esta forma, una palabra que se encuentre de forma más frecuente junto a *excelente* que junto a *horrible* se estimará como una palabra de orientación positiva.

- **Basadas en clasificación de textos**

Una de las técnicas utilizadas con mayor éxito dentro del ámbito de la clasificación temática de grandes colecciones de texto es la basada en Support Vector Machine (SVM). Esta misma técnica es empleada también en el desarrollo de muchos clasificadores emocionales como es el caso de [LeshedyKaye, 2006] donde se presenta un clasificador emocional de blogs que utiliza SVM.

En [Turney y Littman, 2003] se presenta un sistema de identificación de la polaridad del texto basado en Latent Semantic Analysis (LSA). Para saber la polaridad de cada palabra del texto, se calcula la diferencia entre su similitud con un conjunto de palabras positivas y otro de palabras negativas.

El problema fundamental de ambas técnicas radica en el elevado volumen de datos (entrenamiento y test) necesario para asegurar su buen funcionamiento.

- **Basadas en diccionario afectivo**

Estas técnicas se basan en buscar las palabras afectivas que contiene el texto en un diccionario de vocablos afectivos construido previamente. En general, la emoción global del texto se determina a partir de la media de los valores emocionales de cada una de las palabras clave detectadas.

Destaca *Emotional Keyword Spotting* (EKS), debido a su sencillez de implementación. La emoción global del texto se determina a partir de la media de los valores emocionales de cada una de las palabras clave detectadas. Una extensión de EKS es la denominada afinidad léxica, que exporta la emoción de las palabras clave a sus palabras cercanas [Liu, Lieberman, y Selker, 2003]. Ambas son incapaces de detectar cambios de polaridad de la emoción debido a elementos del texto, por ejemplo negaciones [Francisco y Gervás, 2006].

- **Otras técnicas**

Se han estudiado multitud de técnicas para afrontar los problemas de *Sentiment Analysis* con las cuales se han obtenido unos resultados aceptables.

Pueden apreciarse, por ejemplo, los resultados favorables que se obtuvieron al usar una técnica que primero trabajaba sobre la subjetividad de las oraciones de un texto, etiquetándolas, para posteriormente determinar automáticamente la polaridad de las opiniones. Para ellos se utilizaron métodos de clasificación probabilísticas basados en sistemas Bayesianos, Máquinas de Soporte vectorial y modelos de Máxima Entropía.

Caso aparte es el trabajo de [Liu, Lieberman, y Selker, 2003], donde se extraen conceptos de una voluminosa base de conocimiento del sentido común. La ventaja que aporta este sistema es la capacidad de detectar emociones en frases donde a priori no hay una emoción definida explícitamente. Se trata de una técnica compleja debido al tratamiento semántico que debe hacerse de los elementos de la base de conocimiento. Asimismo, [Ovesdotter, Roth, y Sproat, 2005] presentan un sistema complejo que

incorpora técnicas de inteligencia artificial para predecir la emoción del texto en el ámbito de la lectura de cuentos. Éste utiliza, además de palabras afectivas, parámetros del texto como la temática, la longitud de las frases, etc.

Por otro lado, algunas técnicas utilizadas actualmente trabajan primero sobre las frases, formando estimaciones de un conjunto de oraciones, y posteriormente se consideran los párrafos completos dentro del texto, etiquetándolos según las polaridades buscadas [Soo-Min Kim y Eduard Hovy, 2006]. Cuando existe ambigüedad en las polaridades identificadas, estas técnicas utilizan modelos de abstracción conceptuales que permiten evaluar y determinar la mejor polaridad candidata a partir de las oraciones del texto analizado. En general, el enfoque para determinar la polaridad total de un documento a partir de la de sus componentes está basado en un tipo de suma semántica ponderada de los sentimientos individuales presentes.

Otro de los enfoques utilizados en la detección de sentimientos se basa en la identificación de frases comparativas en un texto. El supuesto bajo el que se trabaja es que para expresar los sentimientos hacia un producto, servicio, etc. un usuario/cliente muchas veces lo compara con otro. Para cada frase, se encuentran todas las reglas de detección de polaridad que se satisfacen, y se escoge aquella con la confianza más alta para clasificarla. Si la clase de esta regla es *comparativa* la frase queda clasificada, en caso contrario, esta se define como una frase poco comparativa.

2.2.4. CARACTERÍSTICAS DE IMPORTANCIA PARA LA CLASIFICACIÓN AFECTIVA DE TEXTOS

Convertir un fragmento de texto en un vector de características, o en otra representación que haga que sus características más representativas estén disponibles, es una parte importante en el procesamiento de textos. Existe una extensa línea de investigación dirigida hacia la selección de características para el aprendizaje automático en general. Este capítulo se centra en los hallazgos en la ingeniería de características que son específicos para el análisis afectivo de textos.

2.2.4.1 Presencia de términos frente a Frecuencia de términos

Es habitual en la extracción de información representar un fragmento de texto como un vector de características donde las entradas corresponden a términos individuales. Un hallazgo influyente en el área del análisis de sentimientos es el siguiente. La frecuencia de los términos ha sido tradicionalmente importante en la recuperación de Información (IR) estándar; pero, por el contrario, se ha obtenido [Bo Pang y Lillian Lee, 2002] mejor rendimiento usando la presencia en lugar de la frecuencia. Esto es, vectores binarios de características donde las entradas indican si un término aparece (valor =1) o no (valor =0) formaron una base más efectiva, para la clasificación de la polaridad de críticas, que vectores de características donde las entradas se incrementaban con la frecuencia de aparición de los términos. Este hallazgo puede indicar una importante diferencia entre la clasificación de polaridad y la categorización típica de textos basada en temas. Mientras que un tema es más probable que sea enfatizado por la aparición frecuente de ciertas palabras clave, el sentimiento en general no suele ser realzado a través del uso repetido de los mismos términos.

2.2.4.2 Categorías gramaticales

La información de las categorías gramaticales se ha explotado comúnmente en el análisis de sentimiento y en la minería de opinión. El etiquetado gramatical se puede considerar como una forma ordinaria de desambiguación del sentido de las palabras.

Los adjetivos han sido usados como características relevantes por un gran número de investigadores. Una de las primeras propuestas para predicciones dirigidas por datos de la orientación semántica de las palabras fue desarrollada para adjetivos. Posteriores investigaciones en la detección de subjetividad revelaron una alta correlación entre la presencia de adjetivos y la subjetividad en una frase. Este hallazgo se ha usado para evidenciar que ciertos adjetivos son buenos indicadores de sentimiento, y en ocasiones se ha usado para guiar la selección de características para la clasificación de sentimientos, donde un número de aproximaciones se centran en la presencia o polaridad de los adjetivos para tratar de decidir la subjetividad o la polaridad de unidades textuales, sobre todo en escenarios sin supervisión. En vez de centrarse en adjetivos aislados, también se ha propuesto [Peter Turney, 2002] la detección de sentimientos en documentos basándose en una selección de frases, donde las frases se seleccionaban a través de patrones gramaticales previamente especificados, los cuales solían incluir un adjetivo o un adverbio.

El hecho que los adjetivos sean buenos vaticinadores de que una frase sea subjetiva no implica que otras categorías gramaticales no contribuyan a identificar expresiones de opinión o sentimiento. De hecho, en un estudio sobre la clasificación de la polaridad de críticas cinematográficas [Boo Pang y Lillian Lee, 2002], donde usaba tan sólo los adjetivos como características, se demostró que obtenía peor rendimiento que usando el mismo número de otras unidades gramaticales más frecuentes. Los investigadores resaltaron que los sustantivos (por ejemplo: furia) y los verbos (por ejemplo: querer) pueden ser indicadores fuertes del sentimiento.

2.2.4.3 Sintaxis

Han existido numerosos intentos de incorporar las relaciones sintácticas al conjunto de características a tener en cuenta para el correcto análisis afectivo de documentos. Un análisis lingüístico más profundo parece tener importancia en segmentos cortos de texto.

El análisis sintáctico del texto, puede servir como una base para los modelados de Modificadores de Valencia, como la negación (“no me gusta esta película”), los intensificadores (“la calidad es muy buena”) y los reductores (“el precio es poco competitivo”). Colocaciones lingüísticas y patrones sintácticos más complejos (“el producto es el mejor del mundo”) también ofrecen utilidad para la detección de subjetividad.

2.2.4.4 Negación

El manejo de la negación puede ser un asunto de importancia en el análisis de opiniones y en la detección del sentimiento. Mientras que las representaciones basadas en “sacos de palabras” de “me gusta este libro” y “no me gusta este libro” se consideran muy similares usando para su análisis medidas de similitud comunes, el único término de distinción, el de negación, fuerza a las frases a pertenecer a clases opuestas. No existe una situación paralela en la extracción de información clásica, donde un sólo término de negación juegue un papel tan importante en la clasificación.

Es posible tratar las negaciones directamente como características de segundo orden de un segmento de texto, esto es, donde una representación inicial, como pudiera ser un vector de características, ignoraría la negación, pero donde esa representación se convertiría en otra que sí tuviera en cuenta las negaciones. De forma alternativa, la negación se puede codificar directamente en las definiciones de las características iniciales. Por ejemplo, añadir “NO” a las palabras cercanas al término de negación, por lo que en frases del estilo “no me gustan los plazos”, la característica “gustan” pasaría a ser “gustan-NO”.

Otra dificultad del modelado de la negación es que la negación se puede expresar de formas muy sutiles y, por ejemplo, el sarcasmo y la ironía pueden ser bastante difíciles de detectar

2.2.5. FACTORES QUE DIFICULTAN LA MINERÍA DE TEXTOS DE OPINIÓN

2.2.5.1 Consideraciones del dominio

La precisión de la clasificación basada en sentimiento puede estar influida por el dominio de los artículos sobre los que se está trabajando. Una de las razones es que la misma frase puede indicar un sentimiento diferente bajo el contexto de distintos campos o dominios: Se considera, por ejemplo, la frase "corre y lee el libro" que podría indicar un sentimiento positivo en el ámbito de las críticas de libros, pero un sentimiento negativo para críticas de cine; el adjetivo "imprevisible" es una descripción positiva para el argumento de una película pero una descripción negativa si se habla de la capacidad de dirección de un coche. Diferencias en el vocabulario a través de diferentes campos además aumentan la dificultad cuando se aplica a clasificadores entrenados con datos etiquetados en un dominio para probar datos en otro dominio distinto.

En general, la subjetividad y las opiniones son muy sensibles al contexto, y a modo más general, totalmente dependientes del campo o dominio que se trate (a pesar de que la noción general de opiniones positivas y negativas es bastante constante en diferentes dominios). Nótese que aunque la dependencia del dominio es en parte consecuencia de los cambios en el vocabulario empleado, incluso exactamente la misma expresión puede indicar diferente sentimiento en dominios o campos distintos.

2.2.5.2 Consideraciones del lenguaje

Comparado con los temas, los sentimientos u opiniones pueden ser expresados a menudo de unas formas mucho más sutiles, haciendo que éstos sean difíciles de detectar cuando se considera una frase o algunos términos del documento de forma aislada.

Se considera el siguiente ejemplo:

"Los libros de Jane Austen me exasperan tanto que no puedo ocultar mi furia al lector".

La presencia de palabras como "*exasperar*" y "*furia*" sugieren sentimiento negativo. Entonces, se podría pensar que ésta es una tarea fácil y suponer que la polaridad de las opiniones puede ser identificada, en general, a través de un conjunto de palabras clave.

El problema es que encontrar la colección adecuada de palabras clave puede ser menos trivial de lo que uno inicialmente pudiera pensar y en ocasiones no es suficiente para identificar la polaridad de una opinión.

Se consideran otros ejemplos:

"Si estás leyendo esto porque éste es tu preciado perfume, por favor llévalo sólo cuando estés en casa y mantén las ventanas cerradas" .

"Los libros de Jane Austen me exasperan tanto que no puedo ocultar al lector mi furia. Cada vez que leo "Orgullo y Prejuicio" quisiera desenterrarla y golpear su cráneo con su propio hueso de la espinilla".

Se puede observar que en el primer fragmento no aparecen palabras claramente negativas en la opinión y en el segundo fragmento, aunque la segunda frase indica una opinión extremadamente fuerte, es difícil asociar la presencia de esta dura opinión con palabras clave específicas o frases en esta oración. Es más, la detección de subjetividad puede ser una difícil tarea en sí misma. No solo no se puede identificar con facilidad simples palabras clave con la presencia de subjetividad, sino que también se pueden encontrar patrones, como "el hecho de que", que no garantizan necesariamente la idea objetiva de lo que les sigue. Se puede vislumbrar como la extracción de información orientada a la opinión puede ser realmente una tarea difícil.

2.2.5.3 Consideraciones del tema

Incluso cuando se trata con documentos que pertenecen al mismo dominio, existe todavía una importante y relacionada fuente de variación: el tema del documento. Es cierto que en ocasiones el tema está previamente determinado, como en el caso de las respuestas libres a cuestiones realizadas para una encuesta (como por ejemplo las respuestas vertidas acerca de las prestaciones de un coche concreto). Sin

embargo, en muchas aplicaciones de análisis afectivo, el tema es una característica importante a tener en cuenta.

Una propuesta para integrar sentimiento y tema cuando uno están buscando documentos de opinión acerca de un tema concreto especificado por el usuario, consiste en simplificar primero, desarrollando un análisis de pasada, supóngase por tema, y entonces analizar los resultados obtenidos con respecto al sentimiento [Fabrizio Sebastiani, 2002]. Alternativamente, se puede modelar conjuntamente tema y sentimiento de forma simultánea, o tratar uno antes del otro.

Pero incluso en el caso de estar trabajando con documentos cuyo tema es conocido, no todas las frases dentro de esos documentos están necesariamente englobadas dentro del tema. Para tener en cuenta este aspecto, algunos estudios se basan en desarrollar un proceso en dos pasos; de esta forma cada frase contenida en el documento es primero etiquetada como perteneciente o no al tema en cuestión y el análisis basado en sentimiento sólo se aplica a aquellas frases que han sido encontradas como pertenecientes al tema deseado. Estos trabajos se basan en un supuesto que indica que si una frase ha sido catalogada como perteneciente al tema de interés y manifiesta una polaridad de sentimiento, entonces la polaridad ha sido expresada con respecto al tema en cuestión [Tetsuya Nasukawa and Jeonghee Yi, 2003].

Un asunto relacionado es que es también es posible que un único documento contenga múltiples temas. Por ejemplo, una opinión puede ser una comparación de dos productos. O, incluso cuando un único tema es discutido en un texto, uno puede considerar características o aspectos del producto que representan múltiples (sub) temas. Si se pueden ignorar todos menos el tema principal, entonces una posibilidad es la siguiente: sencillamente considerar el sentimiento global detectado dentro del documento (a pesar del hecho de que éste puede estar formado por una mezcla de opiniones sobre diferentes temas) para ser asociado como tema principal, dejando el sentimiento hacia otros temas sin determinar (es más, esos otros temas puede que nunca sean identificados). Pero es más común tratar de identificar los temas y entonces determinarlas opiniones respecto a cada uno de ellos por separado. En algunos trabajos, los temas importantes son definidos previamente, haciendo esta tarea más sencilla [Jeonghee Yi, Tetsuya Nasukawa, 2003].

2.2.5.4 Incorporación de la estructura del discurso

Comparado con el caso de la tarea tradicional de acceso a la información basada en tema, la estructura del discurso (por ejemplo giros inesperados en documentos) tiende a tener mayor efecto sobre las etiquetas globales de sentimiento. Se ha observado que algunas formas de modelado de la estructura del discurso pueden ayudar a extraer la etiqueta correcta.

No hacen falta escritores experimentados o profesionales del periodismo para producir textos que son difíciles para que las máquinas los analicen. Los escritos de los usuarios de la Web pueden ser tan exigentes, como sutiles, cada uno a su manera. Puede ser más útil aprender a reconocer las características de una opinión. Aún así, es interesante observar la importancia de modelar la estructura de una disertación. Mientras que el tema general de un documento suele ser en el que están centrados la mayoría de los contenidos del mismo, a pesar del orden en el que los diferentes temas son presentados, para las opiniones el orden en el cual las diferentes opiniones son presentadas puede hacer que la polaridad del sentimiento general resultante sea la contraria.

En realidad, al contrario que en la categorización de textos basadas en el tema, los efectos del orden pueden estar muy por encima de los efectos de la frecuencia. Se considera el siguiente fragmento extraído de una crítica de cine:

> *"La película debería ser brillante. Parece un gran argumento, los actores principales son de primera línea, el reparto es bueno también y Stallone está intentando proporcionar una buena interpretación. Sin embargo, no se puede soportar".*

Como se puede ver, las palabras que sugieren una orientación positiva dominan en este fragmento, y aún así, la opinión general es negativa debido a la sentenciadora última frase; mientras que en una clasificación tradicional de textos, si en un documento aparece "coche" con relativa frecuencia, entonces el texto es más probable que trate o que al menos esté relacionado de alguna forma con el tema de los coches.

Se considera también el siguiente ejemplo:

"Odio a las Spice Girls... El porqué ví esta película es verdaderamente una larga historia, pero lo hice, y uno puede pensar que yo desprecié cada minuto de ella. Pero... Vale, estoy realmente avergonzado de ello, pero disfruté. Quiero decir, reconozco que es una película realmente malísima,... (ellos) actúan fatal, una película barata. El argumento es como un caos terrible. Pero me encantó".

A pesar de que predomina el número de sentencias negativas, el sentimiento global sobre la película bajo discusión es positivo, en gran parte debido al orden en el cual esas frases aparecen. Ni que decir tiene, que dicha información se perdería en una representación de todas las palabras sueltas.

La dependencia del orden también se manifiesta en sí misma en muchos pequeños niveles de análisis: "*A es mejor que B*" conlleva exactamente la opinión contraria que "*B es mejor que A*". En general, el modelado de información secuencial y la estructura del discurso parecen más cruciales en el análisis afectivo.

Se ha observado que la posición tiene importancia en el contexto de resumir el sentimiento en un documento [Bo Pang and Lillian Lee, 2004]. En particular, en contraste con el resumen de textos basado en tema, donde el comienzo de los artículos suele servir como una base fuerte en términos de resumir la información objetiva contenida en ellos, las últimas N sentencias de una crítica han demostrado servir mucho mejor para resumir el sentimiento global de un documento más que las N primeras líneas, y ser casi tan bueno como usar (detectadas automáticamente) las N frases más subjetivas, en términos de con cuánta precisión representan el sentimiento global del documento.

2.2.6. TAREAS ENGLOBADAS DENTRO DEL ANÁLISIS DE OPINIÓN

Dentro de la minería de opinión, se pueden identificar varias tareas, todas ellas relacionadas con el etiquetado de un documento dado de acuerdo a la opinión expresada en él.

2.2.6.1 Polaridad de la opinión y grados de positividad

Existen colecciones de problemas que comparten el siguiente caso general: dado un fragmento de texto de opinión, donde se asume que la opinión global es sobre un único tema, clasificar la opinión como dentro de una de dos polaridades de sentimiento, o localizar su posición dentro de una escala continua entre esas dos polaridades. Una gran cantidad de trabajos relacionados con la clasificación de opiniones caen dentro de esta categoría. Se puede señalar que la polaridad o las etiquetas de positividad asignadas pueden usarse simplemente para resumir el contenido de textos de opinión sobre un tema, donde son positivos o negativos, o sólo para recuperar asuntos de una determinada orientación de opinión.

La tarea de clasificación binaria de etiquetar documentos que expresan opiniones generales tanto positivas como negativas, se denomina *clasificación de polaridad del sentimiento* (sentiment polarity classification) o clasificación de polaridad (polarity classification).

Muchos trabajos en la clasificación de polaridad han sido llevados a cabo en el contexto de críticas. Mientras que en este contexto opiniones “positivas” o “negativas” están a menudo basadas en una valoración (por ejemplo ”gusta” o “no gusta”), hay otros contextos donde la interpretación de “positivo” y “negativo” es sutilmente diferente. Un ejemplo es determinar si un discurso político está a favor o en contra del tema bajo debate; una tarea relacionada es clasificar las opiniones de predicciones en un foro sobre las elecciones entre “candidato para ganar” y “no candidato para ganar”. Puesto que estos problemas están todos relacionados con dos clases subjetivas

opuestas, como tareas de aprendizaje automático a menudo son susceptibles a emplear técnicas similares para resolverlos.

Las entradas a un clasificador no son necesariamente siempre opiniones estrictamente hablando. Clasificar un artículo como buena o mala noticia ha sido considerado una tarea de clasificación en muchas investigaciones. Pero una noticia puede ser buena o mala sin ser subjetiva (por ejemplo sin ser reflejo de los estados privados del autor): por ejemplo “el precio del mercado creció” es información objetiva generalmente considerada ser una buena noticia en un contexto apropiado. Aunque no se pretenda proporcionar una definición clara de cuáles deberían ser considerados problemas de clasificación de polaridad, es útil quizás para señalar que (a) en la determinación de la polaridad de una opinión o textos de opinión donde los autores expresen explícitamente sus sentimientos a través de declaraciones como “este portátil es fenomenal”, (podría decirse que) la información objetiva como “batería de larga duración” es a menudo usada para ayudar a determinar la opinión general; (b) la tarea de determinar si un fragmento de información objetiva es bueno o malo no es completamente igual que clasificarlo dentro de una de varias clases basadas en el tema, y por lo tanto hereda los retos involucrados en el análisis de opinión; y (c) la distinción entre información objetiva y subjetiva puede ser sutil. ¿Es “batería de larga duración” objetivo? Además considerar la diferencia entre “la batería dura 2 horas” frente a “la batería sólo dura 2 horas”.

Otro problema general es el *rating inference*, donde se debe determinar la evaluación del autor con respecto a una escala multipunto (por ejemplo de una a cinco “estrellas” para una crítica). Éste puede verse como un problema de categorización de texto en varias clases. Predecir el grado de positividad proporciona una clasificación de la información mucho más precisa y, al mismo tiempo, es un interesante problema de aprendizaje en sí mismo.

Pero a diferencia de muchos problemas de clasificación multi-clase basados en el tema, la clasificación multi-clase relativa a opiniones puede también formularse naturalmente como un problema de regresión ya que la semántica de cada clase puede no corresponderse de forma directa con un punto de la escala considerada. En concreto, cada clase puede tener su propio vocabulario diferente. Por ejemplo, si se está clasificando la evaluación de un autor en tres clases (positivo, negativo y neutral), una

opinión general neutral puede contener una mezcla de lenguaje positivo y negativo. Esto representa oportunidades interesantes para explorar las relaciones entre clases.

2.2.6.2 Detección de subjetividad e identificación de la opinión

Determinar la subjetividad de un documento implica decidir si la naturaleza de un texto dado atiende a hechos (describen una situación dada o un hecho, sin expresar una opinión positiva o negativa en él) o expresa una opinión. Esto puede equivaler a realizar una categorización binaria del texto bajos las categorías Objetiva y Subjetiva.

Los sistemas de extracción de información son propensos a obtener resultados erróneos en aquellos casos en los que las frases contienen un lenguaje subjetivo (opiniones, emociones, etc.). Por ejemplo, los sistemas de extracción de información pueden ser engañados fácilmente con la presencia de lenguaje colorido que contenga, por ejemplo, metáforas o hipérboles. Es por tanto importante el análisis de subjetividad para mejorar la precisión de la información obtenida por los sistemas de extracción.

El trabajo llevado a cabo en la clasificación por polaridad suele asumir que los documentos recibidos son de opinión. Sin embargo, para muchas aplicaciones se puede necesitar decidir si un documento dado contiene información subjetiva o no, o identificar qué porciones del documento son subjetivas.

Varios proyectos evidencian que el problema de distinguir subjetividad frente a instancias objetivas ha resultado ser más complicado que la clasificación de polaridad, y las mejoras en la clasificación subjetiva prometen tener un impacto positivo en la clasificación del sentimiento.

2.2.6.3 Unión del análisis basado en tema y basado en sentimiento

Un supuesto que simplifica la clasificación de sentimiento a nivel de documento es la asunción de que, cada documento que se considere, está enfocado en la materia que concierne. Esto se debe a que se puede asumir que el conjunto de documentos se

creó inicialmente recogiendo documentos que fueran exclusivamente de un tema (por ejemplo, primero se ejecuta un mecanismo de búsqueda de documentos que tratan acerca de un tema en concreto). Sin embargo, es posible que existan interacciones entre tema y opinión que hacen deseable que se consideren ambos de forma simultánea.

Además, un documento que contenga una opinión relevante también puede contener a su vez segmentos alejados del tema, que no tengan interés para el usuario, por lo que se puede desear descartar estos segmentos.

Otro caso de interés, es cuando un documento contiene materia sobre múltiples asuntos que puedan ser de interés para el usuario. En estos casos, es útil que se identifiquen los temas para separar las opiniones asociadas a cada uno de ellos. Dos ejemplos de tipos de documentos para los que este tipo de análisis es apropiado son (1) estudios de comparación de productos y (2) textos que discuten sobre varias características, aspectos o atributos.

2.2.6.4 Puntos de vista y perspectivas

Muchos trabajos en el análisis del sentimiento y las opiniones en textos de orientación política se centran en actitudes generales expresadas a través de textos que no están necesariamente dirigidos hacia un asunto en particular o un tema definido. Por ejemplo, se experimentó [Gregory Grefenstette et al., 2004] la determinación de la orientación política de páginas Web, esencialmente clasificando la concatenación de todos los documentos hallados en dichas webs. Este tipo de trabajo se agrupa bajo el título “Puntos de Vista y Perspectivas”, e incluye la clasificación de textos como liberales, conservadores, libertarios, etc., sitúa los textos de acuerdo a una escala ideológica.

A pesar de que puede usarse la clasificación binaria o multi-clase, las clases no corresponden típicamente a opiniones sobre un sólo tema, o que se puedan definir de forma estricta, sino a una colección de actitudes y creencias.

Otra diferencia con el problema de la clasificación de polaridad es que las etiquetas que se consideran tratan más sobre actitudes que no tienen una correspondencia natural con grados de positividad. Mientras que la asignación de etiquetas sigue siendo un problema de clasificación, si se intentan ofrecer opiniones más

expresivas y abiertas al usuario, ahora se necesitarían resolver los problemas de extracción.

2.2.6.5 Otra información textual no basada en hechos

Algunos investigadores han considerado varios tipos de afecto, emociones como ira, repugnancia, temor, felicidad, tristeza y sorpresa. Una aplicación interesante se encuentra en la interacción humano-máquina: si un sistema determina que el usuario se encuentra disgustado o enfadado, por ejemplo, podría conmutar a un modo de interacción diferente.

Otras áreas de investigación relacionadas incluyen aproximaciones computacionales para el reconocimiento y generación del humor. Muchos aspectos afectivos interesantes en textos como la felicidad o el humor, también se están explorando bajo el contexto de recursos textuales informales como los webblogs.

También se han llevado a cabo investigaciones que se concentran en la clasificación de documentos de acuerdo a su *fuente* o al *estilo de la fuente*, usando la variación estilística detectada de forma estadística como una entrada importante. La identificación de la autoría, es posiblemente el mejor ejemplo.

Otro problema que se ha considerado en escenarios inteligentes y de seguridad, es la detección de lenguaje engañoso.

Muchas aplicaciones podrían beneficiarse de ser capaces también de determinar no sólo si algo es una opinión sino cuál es la intensidad de esa opinión. Los analistas de información necesitan reconocer cambios a lo largo del tiempo en la virulencia expresada por personas o grupos de interés, y detectar cuando la retórica se está acalorando o se está calmando.

Mientras muchas investigaciones se centran en la distinción entre opiniones negativas y positivas u objetivas y subjetivas, es interesante tener en cuenta también la tarea de clasificar la fuerza de las opiniones y las emociones expresadas, considerando distintos niveles de intensidad.

3. ARQUITECTURA DEL SISTEMA

3.1. INTRODUCCIÓN

La arquitectura a desarrollar para obtener un clasificador adecuado estará basada en un sistema en cascada que, tras pasar por una fase de preprocesado de las opiniones y entrenamiento del sistema, debe proporcionar la capacidad de clasificar automáticamente nuevos documentos.

Este enfoque, por tanto, se centra en tener un conjunto de documentos de entrenamiento previamente clasificados, que se usarán para aprender a clasificar nuevos documentos. Para ello, se deben transformar los documentos de su formato inicial a una representación que pueda ser usada por un algoritmo de aprendizaje para la clasificación.

No todos los elementos que aparecen en los documentos serán útiles para su clasificación, es decir, hay elementos que por sí mismos no dicen nada del contenido del documento en el que se encuentran y que por lo tanto pueden ser eliminados; entre ellos se incluyen por ejemplo los signos de puntuación; también aparecen palabras de uso muy frecuente, palabras que aparecen en una gran cantidad de documentos, lo que hace que su poder discriminatorio sea muy bajo; a este tipo de palabras se les conoce como palabras vacías (*stopwords*), ejemplos de ellas son los artículos, pronombres, preposiciones, conjunciones, entre otras semejantes.

Una vez que se ha obtenido una representación adecuada de todos los documentos de entrenamiento (por ejemplo de acuerdo a un modelo de espacio vectorial donde se asigna un peso a las palabras que los conforman) se puede aplicar un método de clasificación (de entre los muchos posibles) para clasificar nuevos elementos.

3.2. FASES DEL DISEÑO DE UN CLASIFICADOR DE TEXTOS

Las fases imprescindibles en las que puede dividirse, de forma genérica, el diseño de un clasificador supervisado de textos son las siguientes (**Figura 7**):

Preprocesado: adquisición de los documentos y obtención de una representación adecuada de los mismos para poder ser empleados por el clasificador. En general, se emplea el modelo de espacio vectorial para la representación.

Reducción de dimensiones: en la clasificación de textos, la elevada dimensión de los vectores puede ser problemática. Por ello, previa a la clasificación, es necesaria una reducción de las dimensiones de los documentos.

Asignación de pesos: cada posición del vector representa una palabra o término dentro del documento y recibe un valor numérico que equivale a la importancia que tiene dicho término dentro del documento.

Entrenamiento: tratamiento de los documentos previamente clasificados para desarrollar un modelo de clasificación mediante un algoritmo de aprendizaje que permita clasificar posteriormente nuevos textos.

Clasificación*:* en función del conocimiento adquirido durante la fase de entrenamiento acerca de las características de cada una de las categorías, el clasificador será capaz de asignar una o varias categorías a nuevos documentos.

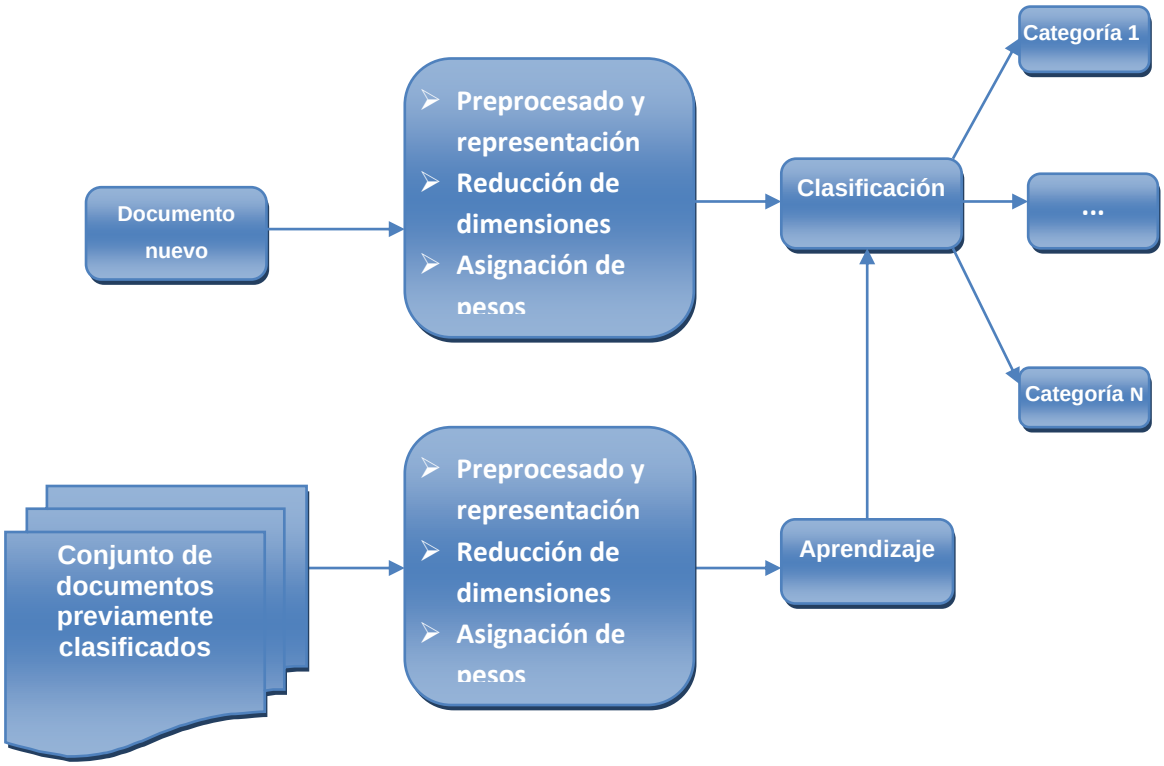

Figura 7. Fases de un clasificador supervisado de textos

3.2.1. PREPROCESADO

La construcción del clasificador automático presenta como punto de partida la obtención del conjunto de documentos o *corpus*. El esquema de aprendizaje empleado se apoya en la existencia de un conjunto inicial de documentos, que han sido previamente clasificados bajo el mismo conjunto de categorías. Una vez que el clasificador ha sido realizado, el resto de documentos formarán el conjunto de validación. Cada documento de este conjunto será introducido en el clasificador y una vez obtenido el resultado se podrá medir el grado de efectividad del clasificador.

Puesto que los documentos no pueden ser interpretados directamente por el clasificador, la obtención de una representación compacta de cualquier documento d_j es necesaria tanto para la fase de entrenamiento como para la de test. Este proceso se conoce como *indexado*. La elección del tipo de representación depende de lo que el usuario considere como la unidad mínima de información y las posibles reglas

lingüísticas del idioma del documento que se pueden aplicar para la combinación de estas unidades.

Como ocurre en el caso de la RI la representación más utilizada es el modelo de espacio vectorial: cada documento es representado mediante un vector de pesos, dj = $\{w_{1j}, w_{2j}, \ldots, w_{|T|j}\}$, donde *T* es el conjunto de términos, que aparecen como mínimo una vez en al menos uno de los documentos del conjunto de entrenamiento . Cada posición del vector indica el grado de importancia de cada término para dicho documento.

3.2.2. REDUCCIÓN DE DIMENSIONES

Para poder realizar la clasificación no sólo es necesario obtener el vector de términos, sino que a su vez es fundamental que dicho vector contenga aquellos términos que son más representativos para poder realizar la clasificación con la máxima efectividad posible. Este paso es fundamental para determinar la calidad del clasificador. Hay que seleccionar aquellas palabras clave que aportan significado y descartar aquellas otras que no contribuyen a realizar la distinción entre documentos. Por tanto, es muy conveniente llevar a cabo un proceso de *reducción de dimensión* (*DR, Dimensionality Reduction*), obteniendo así un conjunto de términos reducidos. Gracias a esto, se consigue evitar el sobreentrenamiento (*overfitting*) en el aprendizaje y se aumenta la eficiencia y efectividad del clasificador.

La elección de las palabras clave se realiza descartando aquéllas que aparecen de forma ocasional en el corpus y las que aparecen muy frecuentemente. El motivo es porque si una palabra aparece muy pocas veces en todo el corpus, seguramente se tratará de un error gramatical u otro caso especial que disminuya la calidad del clasificador. Por otra parte, si una palabra aparece de forma habitual en todas las categorías (preposiciones, verbos auxiliares), será demasiado general para ser utilizada para llevar a cabo la clasificación. Los símbolos de puntuación, los números y los caracteres especiales también son descartados, al igual que las palabras que sólo aparecen en un documento o en un número menor que un umbral. También existe la posibilidad de descartar aquellas palabras cuyo número de apariciones se encuentre por debajo de un determinado umbral fijado por el clasificador. Todas estas palabras se consideran palabras vacías.

Dependiendo de cuáles sean los términos resultantes tras la reducción de dimensión, se pueden distinguir dos posibles esquemas: selección de términos y extracción de términos. En el primero de ellos, el conjunto de términos resultantes $|T'|$ es un subconjunto de $|T|$. Bajo esta premisa, una de las funciones que más comúnmente se emplea en los modelos y técnicas de RI es la de la frecuencia de los términos (número de veces que aparecen). Esto implica la necesidad de normalizar dichos términos, de manera que los recuentos de las frecuencias puedan efectuarse de manera adecuada. Dejando de lado la cuestión de las palabras vacías, hay que tener en cuenta las palabras derivadas del mismo lema, a las que cabe atribuir un contenido semántico muy próximo. Las posibles variaciones de los derivados, junto con formas flexionadas, alteraciones en género y número, etc. hacen aconsejable un agrupamiento de tales variantes bajo un único término. Lo contrario produce una dispersión en el cálculo de frecuencias de tales términos, así como la dificultad de comparar documentos. Esta operación se conoce como s*temming* y se emplea para reducir las dimensiones del vector que representa a un documento.

El segundo de los esquemas para realizar la reducción de las dimensiones del vector que representa un determinado documento es la extracción de términos. En este caso el conjunto $|T'|$ está formado por nuevos términos y no es un subconjunto de $|T|$. Debido a los problemas por polisemia, homonimia y sinonimia, los términos del vector original pueden no ser adecuados para su representación. Cualquier método de extracción de términos debe especificar la forma de extraer los nuevos términos a partir de los antiguos y la forma de convertir la representación original en nuevas representaciones basadas en los nuevos términos.

3.2.3. ASIGNACIÓN DE PESOS

Según el modelo de espacio vectorial, un documento puede ser considerado como un vector $d_j = \{w_{1j}, \dots, w_{|T|j}\}$, donde w_{kj} es un valor numérico que expresa la importancia de la palabra k en el documento j. A continuación se detallan cuatro posibles esquemas para determinar el peso de las palabras que conforman un determinado documento:

- **Representación binaria**

En cada posición del vector se indica la presencia (1) o no presencia (0) de una palabra correspondiente a esa posición.

Tabla 1. Ejemplo empleando representación binaria

Documento	w_1	w_2	w_3
d_1	1	0	1
d_2	1	0	0
d_3	0	1	0
d_4	0	1	1
d_5	1	1	1
d_6	0	0	0
d_7	1	1	0

- **Frecuencia de palabra (TF, Term Frequency)**

A cada palabra se le asigna una importancia proporcional al número de veces que aparece en el documento. El factor TF viene dado por:

$$w_{kj} = tf(t_k, d_j) \qquad \textbf{Ecuación 3}$$

donde TF(d, t) es la frecuencia de la palabra t en el documento d.

Tabla 2. Ejemplo empleando frecuencia de la palabra

Documento	w_1	w_2	w_3
d_1	1	2	3
d_2	2	0	4
d_3	4	3	5
d_4	0	5	3
d_5	3	6	2
d_6	1	2	5
d_7	6	1	7

- **Frecuencia inversa del documento (IDF, Inverse Document Frequency)**

Puesto que existen palabras que aparecen en gran cantidad de documentos, su relevancia será mínima y deberán ser eliminadas del vector. Por ello, serán más importantes las palabras que tengan menor presencia en los documentos analizados.

La importancia de cada palabra es inversamente proporcional al número de documentos que la contienen. El factor IDF de la palabra t viene dado por:

$$idf(t_k) = \log\left(\frac{N}{df(t_k)}\right)$$ **Ecuación 4**

donde N es el número total de textos y df(t) es el número de textos que contienen el término t. Mediante este mecanismo, se asignan pesos elevados a aquellas palabras poco frecuentes en los textos y pesos bajos para las palabras comunes:

$$\log\left(\frac{10.000}{10.000}\right) = 0 \qquad \log\left(\frac{10.000}{5.000}\right) = 0.301$$

$$\log\left(\frac{10.000}{20}\right) = 2.698 \qquad \log\left(\frac{10.000}{1}\right) = 4$$

- **TF·IDF**

Es posible obtener mejores prestaciones cuando se combinan los dos mecanismos anteriores, empleando la siguiente expresión:

$$w_{kj} = tf(t_k, d_j) \cdot idf(t_k)$$ **Ecuación 5**

Según esta fórmula, cuanto más aparezca un término en un documento, más representativo será para su contenido, y cuantos más documentos contengan dicho término, menos discriminante será para realizar la clasificación. La importancia que tiene un término para un determinado documento se calcula únicamente en función del número de apariciones. El orden en que aparece en el documento y el papel sintáctico que juega no se tiene en consideración.

Para conseguir que los pesos reciban valores dentro del intervalo [0, 1] y que los vectores tengan igual longitud, se suele emplear lo que se conoce como *normalización del coseno*:

$$w_{kj} = \frac{tf(t_k,d_j)\cdot idf(t_k)}{\sqrt{\sum_{s=1}^{|T|}\left(tf(t_s,d_j)\cdot idf(t_s)\right)^2}}$$ **Ecuación 6**

3.2.4. ENTRENAMIENTO

Para desarrollar el clasificador se debe seleccionar un número apropiado de documentos previamente clasificados a los que se denomina conjunto de entrenamiento. Este conjunto, una vez transformado adecuadamente de su formato inicial a una representación que pueda ser tratada por el algoritmo de aprendizaje, servirá para entrenar el sistema y capacitarle para decidir la clase a la que pertenecerán nuevos documentos entrantes cuya clase se desconoce.

3.2.5. CLASIFICACIÓN

Un sistema de clasificación automática de documentos, consiste, en sentido amplio, en un conjunto de algoritmos, técnicas y sistema capaces de asignar a un documento una o más categorías dentro de una jerarquía dependiendo de cuál sea, por ejemplo, su afinidad temática.

El escenario de aplicación de este proyecto se basa en que las clases y los grupos son determinados previamente por personas, y la labor del sistema es simplemente asignar cada documento a una de esas clases definidas a priori. Este modelo se conoce como clasificación automática supervisada, en el sentido de que requiere la supervisión o intervención humana, tanto para diseñar las clases o categorías como para entrenar el sistema.

Una vez obtenida una representación adecuada de la colección de documentos de entrenamiento, se puede aplicar un método de clasificación automática (de entre los muchos existentes) para clasificar nuevos elementos.

Como se verá en la siguiente sección (**4. Diseño e Implementación del Sistema**) se han elegido para el desarrollo del proyecto dos métodos en concreto para realizar la clasificación, uno de ellos aplicando el algoritmo de los K vecinos más próximos (kNN-k Nearest Neighbour) y otro basado en el empleo de un diccionario afectivo.

4. DISEÑO E IMPLEMENTACIÓN DEL SISTEMA

4.1. INTRODUCCIÓN

De acuerdo a la arquitectura comentada en el apartado anterior, se va a detallar el diseño de los sistemas implementados. Se ha optado por realizar la implementación de dos sistemas que, empleando diferentes métodos de clasificación, permiten, en última instancia, clasificar adecuadamente los textos requeridos, a fin de evaluar y comparar su efectividad.

Siguiendo las indicaciones del modelo de diseño de un clasificador automático explicado en el **apartado 3.2**, la arquitectura de los sistema desarrollados consiste en tres bloques principales: preprocesado, entrenamiento y clasificación. En la **Figura 8** se ilustra de forma general el modelo empleado. Como se puede observar, se trata de un sistema en cascada, donde la salida de cada etapa se convierte en una entrada para la siguiente. En las siguientes secciones se explicará de forma detallada el desarrollo de cada una de las fases para los dos sistemas desarrollados.

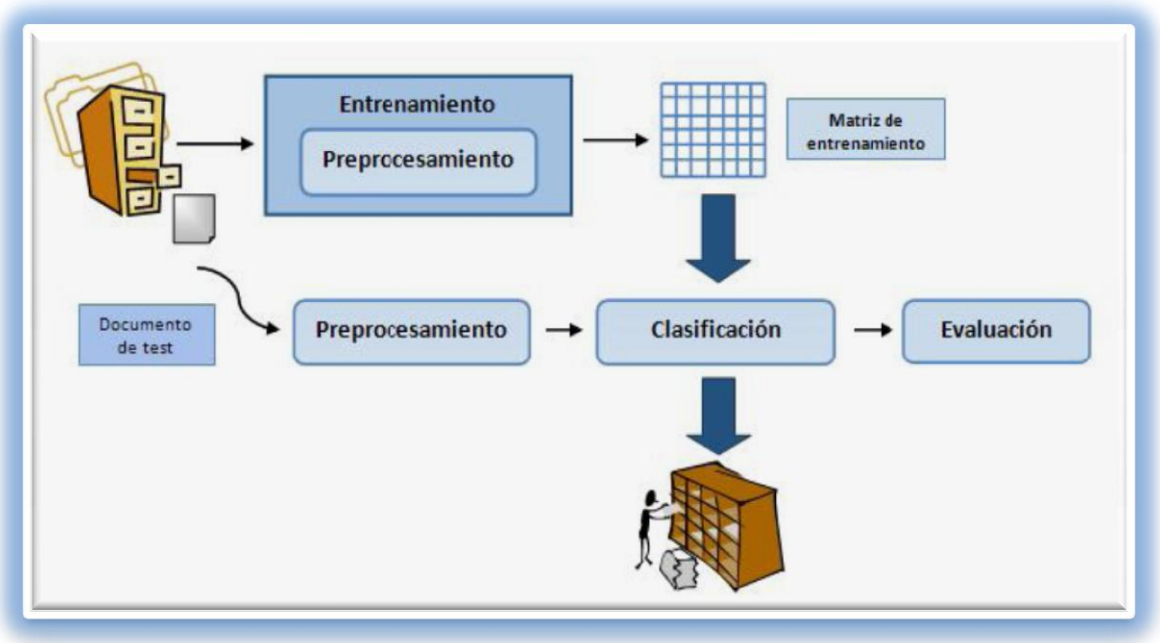

Figura 8. Arquitectura del sistema

Para el desarrollo de un sistema de clasificación automática, existen dos posibles escenarios: Por un lado, en el caso de la clasificación no supervisada, los documentos se clasifican, en función de su contenido sin asistencia manual y sin la existencia de categorías determinadas previamente o cuadros de clasificación establecidos a priori. Por otro lado, el tipo de clasificación que se va a llevar a cabo se trata de una clasificación supervisada, es decir, se diseñará un clasificador que partiendo de una serie de categorías gramaticales diseñadas a priori, se encarga de asignar cada documento a la categoría correspondiente. Requiere la elaboración manual o intelectual del conjunto de categorías. En este escenario, para que el sistema aprenda y sea capaz de asignar posteriormente la clase adecuada es necesaria una fase previa de entrenamiento del clasificador.

En primer lugar es necesario elegir un conjunto de documentos con los que desarrollar el proyecto. En el caso que se está estudiando este conjunto se trata de textos de opinión, más concretamente de **críticas de cine** donde usuarios de Internet expresan su opinión acerca de diversas películas. El motivo de la elección de este corpus es que en las críticas cinematográficas la variedad de opiniones es muy amplia lo cual permite obtener con relativa facilidad muchos documentos distintos que pertenezcan a clases diferentes. Una descripción más detallada del *corpus* empleado se presenta en el **apartado 5.1.**

Una vez elegido el *corpus* se deben obtener dos conjuntos: entrenamiento y test. Por cada una de las categorías que conforman este *corpus* (2, 4 o 5 categorías según los casos de estudio), el 70% de sus ficheros se incluyen como parte del conjunto de entrenamiento y el 30% restante pertenecerán al conjunto de evaluación (esta proporción es usada con frecuencia en minería de datos). Todas las críticas a clasificar pertenecen sólo a una de las categorías contempladas en la fase de entrenamiento.

En el **apartado 4.2** se explicará con detalle el diseño y la implementación de cada una de las fases del primer sistema desarrollado el cual está basado en el algoritmo de los k vecinos más próximos (kNN-K Nearest Neighbour). Del mismo modo, en el **apartado 4.3** se explicará con detenimiento el diseño del clasificador desarrollado basado en el empleo de un *diccionario afectivo*. Las pruebas realizadas y los resultados obtenidos con ambos sistemas se presentarán en el **apartado 5.3.**

4.2. DISEÑO DEL SISTEMA BASADO EN EL ALGORITMO KNN

De acuerdo a lo explicado en el apartado anterior, y en el capítulo 3, donde se describen las fases de un clasificador, se puede dividir el desarrollo de este sistema en 3 etapas básicas: preprocesado, entrenamiento y clasificación.

Tanto los documentos pertenecientes al conjunto de entrenamiento como los pertenecientes al de test son introducidos en una fase de preprocesado, con el objetivo de obtener una representación estructurada de los mismos.

Como resultado de la fase de entrenamiento se obtiene una matriz en la que se encuentran almacenados todos aquellos términos que han sido considerados como relevantes para poder realizar la clasificación y un factor que indica su importancia dentro de cada documento (las técnicas empleadas para obtener esta matriz se explicarán con detalle en los siguientes apartados).

El siguiente módulo consiste en la clasificación propiamente dicha. Las entradas de esta etapa son la matriz de entrenamiento y el documento nuevo a clasificar. Entre los distintos algoritmos de clasificación automática de textos mencionados en el **apartado 2.1.3**, se ha optado, en este caso, por emplear el algoritmo de los *k vecinos más próximos (KNN-K Nearest Neightbour)* debido, en gran medida, a su sencillez y eficiencia. Se trata de un método de clasificación supervisada no paramétrico que ha sido empleado en multitud de aplicaciones dentro del ámbito de la minería de datos como por ejemplo, para el reconocimiento de patrones, procesado de imágenes, etc. Se trata de un algoritmo muy sencillo basado en encontrar los k vecinos más cercanos o parecidos al que se usa para realizar la consulta.

Por último, una vez encontrados los k documentos más parecidos al documento de test, según una medida de distancia como la distancia Euclídea, se procede, en base a esto, a determinar la categoría bajo la cual debe ser clasificado el nuevo documento.

4.2.1. PREPROCESADO

Este módulo es ejecutado tanto en la fase de entrenamiento como en la de test y su objetivo no es otro que el de obtener una representación estructurada y simplificada de los documentos que ayuden a su posterior procesado. Se deben trasformar los documentos de su formato inicial a una representación que pueda ser empleada por un algoritmo de aprendizaje para la clasificación.

4.2.1.1 Análisis morfosintáctico y lematización

Un documento es un escrito que contiene información, estructurada en frases u oraciones. La oración es la mínima unidad del habla con sentido completo y está compuesta por palabras, que son segmentos limitados por pausas o espacios. Estos segmentos pueden ser nombres (comunes o propios), verbos (en sus diversas conjugaciones), adjetivos, adverbios, artículos, preposiciones, etc. Son por tanto las unidades que aportarán sentido a los textos.

No obstante, se debe tener en cuenta que, no todas las palabras son relevantes no todos los elementos que aparecen en los documentos son útiles para su clasificación, es decir hay elementos que por sí mismos no dicen nada del contenido del documento en el que se encuentran y que por lo tanto pueden ser eliminados; entre ellos se incluyen por ejemplo los signos de puntuación. A este tipo de palabras se les conoce como palabras vacías (*stopwords*), ejemplos de ellas son los artículos, pronombres, preposiciones, conjunciones, entre otras semejantes.

En el caso de este sistema de clasificación desarrollado, puesto que se trata de textos de opinión, se ha considerado que los verbos y los nombres son importantes pero sobre todo lo serán los adjetivos y, en última instancia, los adverbios (se ha considerado si emplearlos o no). Otras unidades como artículos, preposiciones, pronombres, signos de puntuación etc. serán descartadas a fin de reducir, en primera instancia, las dimensiones de la colección de entrenamiento y hacerla más manejable, para continuar el proceso de clasificación automática Por tanto, serán éstas (sustantivos, verbos y adjetivos) las unidades a tener en cuenta para representar los documentos. Esta forma de reducción de dimensiones está basada en la extracción de términos (explicada en el **apartado 3.2.2**) puesto que los términos seleccionados pertenecen al conjunto inicial.

El análisis morfológico de los textos proporcionará varios significados gramaticales para cada una de las palabras por lo que ha parecido interesante realizar un proceso de **desambiguación semántica** de las mismas para obtener aquellos significados que de verdad se adapten al contexto.

Por otra parte, muchas palabras, aunque diferentes, tienen la misma raíz léxica por lo que se ha realizado un **proceso de lematización**, con el cual se busca reducir las palabras a su raíz. La lematización es un proceso complejo desde el punto de vista morfosintáctico, donde las inflexiones y las formas varias de una palabra se reducen. Cuando se lematiza un texto, se reemplaza cada palabra del mismo por su lema; un texto que ha sido lematizado, entonces, contendrá todas las formas verbales representados por su infinitivo, todas las formas sustantivas representadas por su forma masculino singular, etc.

Al trabajar con la raíz de la palabra se consigue tener una mejor cobertura del idioma (fundamental en un idioma fuertemente flexionado como el español) y se obtiene una representación única para vocablos que compartan la misma raíz, lo cual reducirá considerablemente las dimensiones del clasificador y proporcionará, además, unos mejores resultados.

Para poder llevar a cabo el análisis necesario para extraer sólo los elementos necesarios del texto, se ha empleado una herramienta de procesamiento lingüístico: ***STILUS Core***, producto distribuido por la empresa *DAEDALUS*. Se trata de una biblioteca software de herramientas para procesamiento lingüístico en castellano: filtrado, segmentación y etiquetado morfosintáctico de textos, análisis sintáctico superficial, desambiguación morfológica, extracción de resúmenes, etc. En la **Figura 9** se muestran las distintas opciones que presenta esta herramienta.

```
mmartin@plato:~$ /home/miracle/tools/stilus-core-es/stilus-core-es -h
STILUS Core (es)
(c) 1998-2006 DAEDALUS
/home/miracle/tools/stilus-core-es/stilus-core-es:
-d ruta     Directorio de datos (/home/ingling/diccio/)
-t ruta     Directorio temporal (/tmp/)
-m tipo     texto,html,tex (defecto: texto)
-r          (sólo texto) El carácter \r separa líneas
-br         (sólo html) La etiqueta BR separa párrafos
-V version  Versión (sólo libros de estilo específicos)
Modos:
-Mv         Versión de los recursos lingüísticos
-Mam        Analisis morfológico
-Mas        Analisis sintáctico
-Mr         Resumidor de texto
-Mgm        Generador morfológico
-Mcv        Conjugador verbal
Opciones del análisis morfológico/sintáctico:
-pd         Detectar palabras desconocidas
-g          Desambiguar análisis
-l          Imprimir descripción completa de etiquetas
-i          Imprimir información semántica completa
Opciones del resumidor de texto:
-nf         Número de frases (3)
-Pn         Puntuacion de los nombres (10)
-Pv         Puntuacion de los verbos (5)
-Pa         Puntuacion de los adjetivos (0)
-Pd         Puntuacion de los adverbios (0)
-Pg         Puntuacion de las negritas (2)
-mp         Mostrar las puntuaciones
mmartin@plato:~$
```

Figura 9. Opciones de *STILUS Core*

Para el diseño del sistema realizado sólo ha sido necesario emplear el análisis morfológico (opción –Mam) que la herramienta proporciona. Con este análisis se determina la forma, clase o categoría gramatical de cada palabra en una oración. Por otra parte también se ha realizado una desambiguación morfosintáctica (opción –g) para filtrar los análisis inválidos en el contexto donde aparecen las palabras.

Se realiza sobre todas las críticas del *corpus* la siguiente instrucción, de manera que se almacenan los análisis morfológicos de cada una de las opiniones en unos nuevos ficheros.

mmartin@plato:~$ cat /home/mmartin/B_opinion.txt | /home/miracle/tools/stilus-core-es/stilus-core-es -d /home/miracle/tools/stilus-core-es -Mam -g > /home/mmartin/analisis_B_opinion.txt

En la **Figura 11** se puede comprobar cuál es el resultado obtenido al aplicar esta herramienta sobre uno de los textos considerados (**Figura 10**):

> *Me gustó esta película, es tierna, romántica y además los actores principales Jonny Deep y Leo Dicaprio están insuperables, está entre mis títulos favoritos. Tiene también tintes de comedia por lo que la sonrisa está asegurada. Os la recomiendo.*

Figura 10. Ejemplo de opinión a analizar

```
Me          0     2     PPMS1-HAN7|yo|me      PPFS1-HAN7|yo|me      PPMS1-HDN7|yo|me
      PPFS1-HDN7|yo|me
gustó       3     5     VI-S3SBL-N5|gustar|gustó
esta        9     4     DDFSNN7|este|esta
película    14    8     NCFS--N-N6|película|película      NCFS--N-N6|película|película
,           22    1     1O--|,|,
es          23    2     VI-S3PIA-N8|ser|es
tierna      26    6     APFS--NN4|tierno|tierna
,           32    1     1O--|,|,
romántica   33    9     NCFS--N-N4|romántico|romántica
y           43    1     CCY--N8|y|y
además      45    6     E-X-N6|además|además
los         52    3     TDMPN9|el|los
actores     56    7     NCMP--N-N5|actor|actores      NCMP--N-N5|actor|actores
principales 64    11    NCMP--N-N6|principal|principales
Jonny       76    5     NPMS--N-N2|Jonny|Jonny
deep        82    4     ?
y           87    1     CCY--N8|y|y
Leo         89    3     VI-S1PTL-N5|leer|leo
Dicaprio    93    8     NPMS--N-N2|DiCaprio|DiCaprio  NPMP--N-N2|DiCaprio|DiCaprio
      NPFS--N-N2|DiCaprio|DiCaprio
están       102   5     VI-P3PIA-N7|estar|están
insuperables 108  12    APMP--DN3|insuperable|insuperables    APMP--
NN3|insuperable|insuperables    APFP--DN3|insuperable|insuperables    APFP--
NN3|insuperable|insuperables
,           120   1     1O--|,|,
está        121   4     VI-S3PIA-N7|estar|está  VM-S2-IA-N7|estar|está  VM-S6OIA-N7|estar|está
entre       126   5     Y-N7|entre|entre
mis         132   3     SDMP1SHT8|mío|mis
títulos     136   7     NCMP--N-N5|título|títulos
favoritos   144   9     APMP--NN4|favorito|favoritos
.           153   1     1O--|.|.
*
Tiene       155   5     VI-S3PTL-N7|tener|tiene
también     161   7     E-X-N7|también|también
tintes      169   6     NCMP--N-N3|tinte|tintes  NCMP--N-N3|tinte|tintes
de          176   2     Y-N9|de|de
comedia     179   7     NCFS--N-N5|comedia|comedia
por_lo_que  187   10    CSVLUN5|por_lo_que|por_lo_que
la          198   2     TDFSN9|el|la
sonrisa     201   7     NCFS--N-N4|sonrisa|sonrisa
está        209   4     VI-S3PIA-N7|estar|está
asegurada   214   9     APFS--DN4|asegurado|asegurada  APFS--NN4|asegurado|asegurada
.           223   1     1O--|.|.
Os          224   2     PPMP2-HAN6|tú|os      PPFP2-HAN6|tú|os      PPMP2-HDN6|tú|os
      PPFP2-HDN6|tú|os      NCMS--N-C3|Os|Os
la          227   2     PPFS3-UAN9|él|la
recomiendo  230   10    VI-S1PTL-N5|recomendar|recomiendo
*
```

Figura 11. Salida de *STILUS Core*

El análisis obtenido para cada palabra del texto consiste en lo siguiente:

- Palabra de entrada
- Desplazamiento respecto al origen del fichero
- Longitud de la palabra
- Separados por tabulaciones:
    - La o las categorías morfológicas que puede tener la palabra, codificada según el etiquetario de *STILUS*.
    - El o los lemas de cada una de las categorías
    - La información semántica de la palabra
    - La forma canónica de la palabra (mayúsculas/minúsculas)

Como se puede observar en la salida se obtiene el lema de cada una de las palabras con lo cual se puede trabajar únicamente con la raíz de las mismas. En el caso del sistema que se está implementando, y teniendo en cuenta las categorías gramaticales de interés, se obtendrán todas las formas verbales representados por su infinitivo, todas las formas sustantivas representados por su forma masculino singular al igual que en el caso de los adjetivos.

4.2.1.2 Búsqueda de sinónimos y modificadores

Además de poder seleccionar sólo aquellas palabras que pertenezcan a las categorías gramaticales de interés, se han elaborado unos archivos auxiliares que serán de ayuda para un mejor tratamiento de las palabras contenidas en los documentos. Estos archivos contienen, en un caso, una serie de sinónimos de distintas palabras cuyo significado es potencialmente afectivo y, en otro caso, una serie de modificadores de la intensidad de las palabras (muy, poco, bastante, etc.).

Al estar trabajando sobre un dominio específico, existe cierto control sobre los términos que le pertenecen, por lo que se ha decidido hacer un emparejamiento entre distintas acepciones que puede tener un mismo concepto a una sola representación, es decir, si a un concepto se le puede llamar de distintas formas, se considera unificarlas y dentro del proceso de clasificación considerarlas bajo una forma única. Para ello se ha elaborado un archivo (**Anexo A**) que contiene multitud de **sinónimos** para englobar bajo

un único término aquellas palabras que habitualmente aparecen en los documentos con los que se trabaja.

Un ejemplo de una de las entradas de este diccionario es el siguiente:

excelente perfecto emocionante maravilloso genial impresionante fantástico espectacular formidable espléndido excepcional magnífico estupendo fabuloso extraordinario superior prodigioso soberbio fascinante tremendo magistral apasionante inolvidable precioso sobresaliente admirable memorable impecable;

Figura 12. Entrada del fichero de *Sinónimos*

De esta forma, si en una crítica aparece, por ejemplo, el adjetivo *excelente* y en otras *fantástico* o *maravilloso* se considerará que los textos contienen la misma palabra, que es la primera de las acepciones de la lista, en este caso sería *excelente*.

Por otra parte se han elaborado también unos archivos (**Anexo B**) que contienen una serie de **modificadores** (de verbos y de adjetivos), es decir, palabras que pueden variar la importancia de otras que son de interés dentro del texto. Un ejemplo de las entradas de estos archivos se muestra a continuación (**Figura 13**), donde se observan algunos modificadores considerados precedidos del valor adicional que introducirán respecto a la frecuencia de aparición de las palabras a las que modifiquen.

2 súper totalmente absolutamente;
1 muy;
0.5 bastante;
-0.5 poco;

Figura 13. Entrada del fichero de *Modificadores*

Una vez detectados en el texto estos modificadores, el valor que tengan asociado en el fichero que los contiene (valores decididos de acuerdo al grado de modificación que introducen) será sumado a la frecuencia de aparición de las palabras (verbos o adjetivos) a las que estén modificando (cuando se realice la asignación de pesos). De esta forma, si delante de la palabra *bueno* se detectara el modificador *muy* se considerará que en ese caso la palabra *bueno* no contribuye con el valor uno a su frecuencia total de aparición sino que equivaldrá a haber aparecido dos veces ya que el modificador de intensidad de la palabra, en este caso, incrementa en uno su valor original.

4.2.2. ENTRENAMIENTO

Tras realizar la fase de preprocesado y haber obtenido los términos que mejor representan a los documentos, se puede pasar a construir la **matriz de entrenamiento** del sistema (basada en el modelo de espacio vectorial) que servirá posteriormente para llevar a cabo la clasificación de nuevos documentos.

4.2.2.1 Modelo de espacio vectorial

La idea básica detrás de este modelo es construir una tabla donde se manejan los documentos y las palabras que estos contienen, asignándoles un peso a cada una de ellas. Cada vector que conforma la matriz representa a un documento y la distribución de las palabras que en él aparecen. Se trata de una matriz de m x n, donde m son los documentos y n las palabras registradas.

Para evitar tener en cuenta cada una de las palabras que aparecen en el conjunto de entrenamiento, lo cual supondría un coste elevado para la construcción de la matriz de entrenamiento y la posterior comparación con los nuevos documentos, se ha elaborado una **lista de palabras** que ayudará a simplificar el sistema.

Esta lista se construye de manera dinámica (mediante un array asociativo) a medida que se realiza el preprocesado de los documentos de entrenamiento. Se añade una nueva palabra a la misma, siempre que sea de una de las categorías gramaticales deseadas (sustantivos, verbos y adjetivos) y siempre que la palabra no haya sido introducida en la lista con anterioridad. En resumen, la lista contiene las palabras de las categorías deseadas presentes en uno o varios documento de entrenamiento y un valor asociado a cada una de ellas que indica el número de documentos diferentes en los que aparece dicha palabra. Esta frecuencia de aparición de las palabras en diferentes textos podría servir para aplicar si se deseara el algoritmo TF*IDF (comentado en el **apartado 3.2.3**) y para hacer una reducción de dimensiones que se explicará con posterioridad.

4.2.2.2 Reducción de la dimensionalidad

No se pueden considerar todas las palabras que integran la colección de entrenamiento dentro del modelo de espacio vectorial, ya que la dimensión que tendría sería enorme. Existen distintas técnicas para reducir la dimensión, como la frecuencia documental, que considera un valor mínimo de apariciones que debe tener cada palabra dentro del total de documentos, para discriminar aquellas cuya aparición sea muy pequeña y dejar las que presenten mayor frecuencia documental. En este caso se aplicará una reducción considerando que las palabras que integrarán la lista, y por tanto la matriz de entrenamiento, serán aquellas cuya presencia en los distintos documentos no sea ni muy baja ni demasiado elevada (podrán ir variándose los umbrales). Esta decisión se debe a que las palabras cuya aparición es muy frecuente, al aparecer en gran cantidad de documentos, tienen un poder discriminatorio muy bajo. Por otro lado, las palabras que son muy específicas de algunos textos y que no aparecen prácticamente en ninguno más, aumentarían innecesariamente las dimensiones de la matriz de entrenamiento y además no ayudarían, en principio, a realizar una adecuada comparación con los nuevos documentos a clasificar.

4.2.2.3 Asignación de pesos

Existen distintos tipos de ponderaciones para las palabras dentro del modelo de espacio vectorial que se ha empleando. De entre los métodos comentados en el **apartado 3.2.3** se han decido considerar para el sistema las siguientes ponderaciones:

- **Asignación booleana:** Es la más sencilla. El peso de la palabra es 0 si no aparece en el documento, y 1 en caso de que aparezca.
- **Asignación basada en TF (Term Frecuency)**: El peso del término depende de la cantidad de ocurrencias que tenga dentro del documento. Posteriormente será normalizado para obtener valores dentro del rango [0,1].

En el caso de la asignación basada en TF (ya que en el caso binario no tiene sentido), es en este punto donde a la frecuencia de aparición de los términos se le sumará la influencia de los modificadores de las palabras detectados, de acuerdo a lo comentado en el apartado anterior.

Una vez finalizada la fase de entrenamiento, se obtiene una matriz que servirá para realizar la comparación de los ficheros de entrenamiento con los nuevos documentos de entrada para una posterior clasificación. Esta matriz es tal que sus filas corresponden a cada uno de los documentos empleados para entrenar el sistema y sus columnas corresponden al peso (binario o basado en TF) de cada palabra de la lista general (construida a partir de los documentos) en cada uno de los ficheros del *set* de entrenamiento. No es necesario almacenar las palabras en sí mismas en la matriz de entrenamiento ya que al disponer de la lista general creada previamente, el orden de la lista (donde sí están almacenadas las palabras) es el orden de las columnas de la matriz y este orden se mantendrá también al constituir los vectores que representarán a los futuros documentos a clasificar.

A continuación puede verse un ejemplo sencillo de la lista que se construiría si sólo se tuvieran dos pequeñas críticas como documentos de entrenamiento y cómo serían los vectores representativos una vez realizada la asignación de pesos (en este caso, aunque en la práctica si se realiza, no se han normalizado los resultados de forma que pueden entenderse con más facilidad los pesos asignados).

4.2.3. CLASIFICACIÓN

Una vez obtenida la representación de la colección de documentos de entrenamiento, se puede aplicar un método de clasificación automática para clasificar nuevos elementos. Por tanto, en este modulo se aborda el problema de categorizar cada uno de los documentos que componen el conjunto de test dentro de una de las categorías definidas.

De entre los diversos tipos de algoritmos de clasificación existentes se ha optado por emplear el de los ***k vecinos más cercanos*** (*kNN, k Nearest Neighbour*) por su sencillez y eficacia. En esta técnica, cada vector de entrada es comparado con los vectores de otros documentos para encontrar los k documentos más cercanos (o más similares) y utiliza las categorías de estos k documentos para determinar la categoría correspondiente del documento de prueba.

Para poder aplicar el algoritmo de clasificación elegido, primero se debe realizar el preprocesado de cada uno de los nuevos documentos (al igual que se hizo con los textos del *set* de entrenamiento) y posteriormente se construye el vector que representa a cada uno de ellos. Para construir este vector representativo, se busca la aparición de cada una de las palabras de la lista general (construida durante el entrenamiento del sistema) en el documento que se desea categorizar. La asignación de pesos puede ser de nuevo binaria o basada en TF, es decir representando la existencia o no de cada uno de los términos en el documento o representando la frecuencia de aparición normalizada de los términos dentro del mismo.

Una vez que se ha obtenido el vector representativo del documento a clasificar bastará con determinar la distancia entre dicho vector y los vectores que componen la matriz de entrenamiento del sistema, para así poder determinar los k documentos más próximos o semejantes.

En este caso, para estimar el grado de similitud entre cada documento del conjunto de entrenamiento y el documento de test se calcula el **coeficiente del coseno**, ampliamente utilizado en sistemas de IR. Se trata de una medida de semejanza entre dos vectores a través del cómputo del ángulo que existe entre ambos de forma que un menor ángulo representa una mayor similitud. La expresión que se emplea es la siguiente:

$$SIM(E_x, T_y) = \frac{\sum_{i=1}^{N} E_i T_i}{\sqrt{\sum_{i=1}^{N} E_i^2 \cdot \sum_{i=1}^{N} T_i^2}}$$ **Ecuación 7**

Donde E representa el vector de un documento de entrenamiento, T es el vector de la consulta, T_i y E_i equivalen al peso asignado al término i dentro de ese documento y N es el número total de términos de la lista.

En la expresión se asume que ambos vectores tienen igual número de términos y en este caso se cumple ya que, tanto los vectores que representan los documentos de entrenamiento como los que representan los de test, tienen una longitud igual al número de palabras almacenadas en la lista general construida durante la fase de entrenamiento. Gracias a esto, se puede reducir el coste computacional del sistema y no

almacenar en los vectores, por ejemplo, palabras de los documentos nuevos a categorizar que no existan en ninguno de los textos de entrenamiento.

Por último, una vez obtenida la distancia entre el documento nuevo a categorizar y cada uno de los documentos del conjunto de entrenamiento, se seleccionarán de estos últimos aquellos k documentos que presenten una similitud mayor con el documento de test que se esté tratando. La categoría final asignada al nuevo documento será aquella que más se repita entre los k documentos de entrenamiento seleccionados. En caso de empate entre las categorías que más se repiten, se ha optado (aunque podría haberse aplicado otro criterio) por asignar aquella categoría que mayor grado de positividad presenta dentro de la escala de categorías definida, es decir que, por ejemplo, ante un empate entre el número de documentos de entrenamiento más perecidos buenos y excelentes la categoría asignada será *excelente.*

Este procedimiento se lleva a cabo con cada uno de los documentos del conjunto de test para así, una vez obtenidos los resultados, comparar las clases reales de los documentos con las clases asignadas por el sistema, lo cual permitirá evaluar las prestaciones del clasificador.

4.2.4. IMPLEMENTACIÓN DEL SISTEMA

El sistema de clasificación descrito ha sido implementado en el lenguaje de programación **PHP** (*PHP Hypertext Pre-processor*), ampliamente utilizado por su sencillez, versatilidad y potencia. Estos han sido los motivos que han llevado a la utilización de este lenguaje unido a que no necesita compilación, al tratarse de un lenguaje interpretado, y que es de acceso libre y gratuito.

En las siguientes secciones se explicará de forma detallada el contenido de los ficheros que se procesan en cada uno de los módulos y, a grosso modo, las funciones implementadas para poder llevar a cabo el diseño del clasificador.

El diagrama general del sistema se puede observar en la siguiente figura.

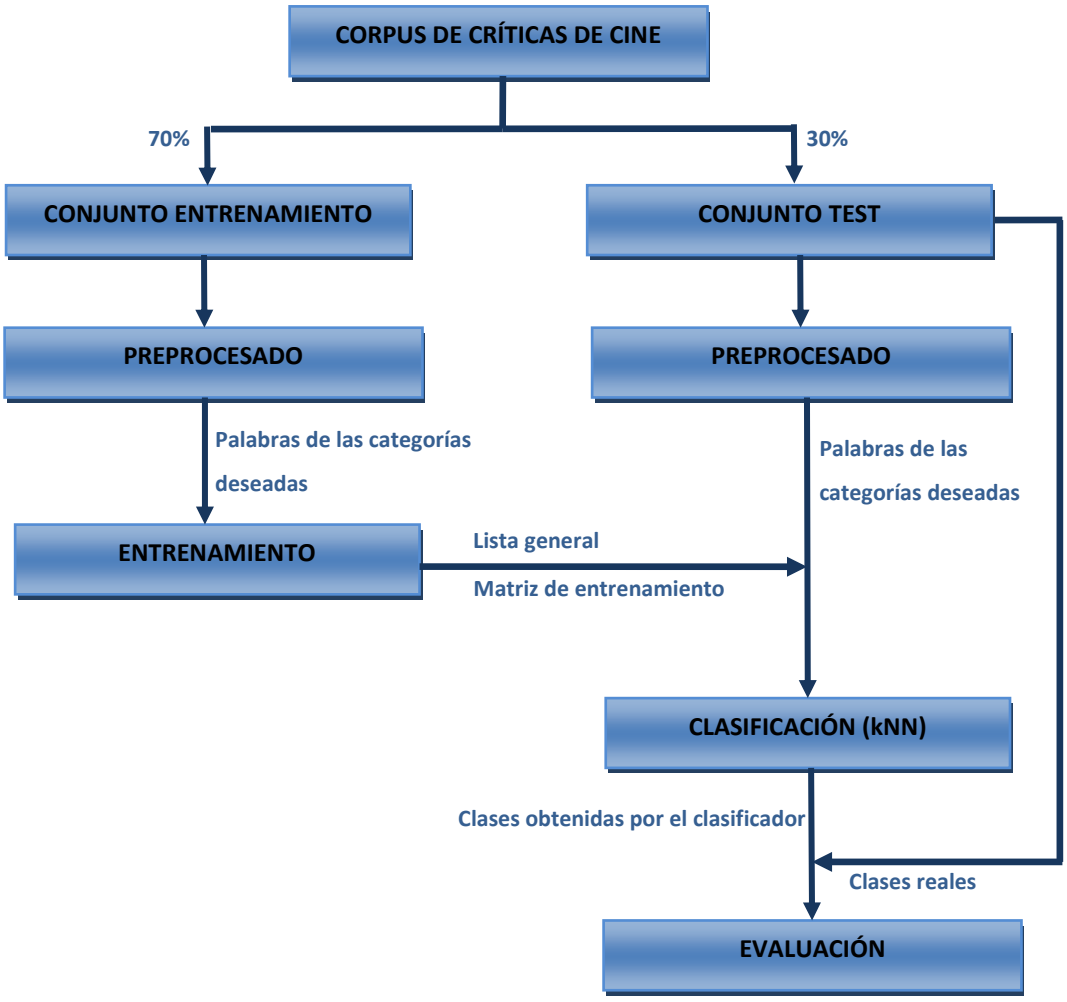

Figura 14. Diagrama del sistema basado en kNN

4.2.4.1 Conjunto de entrenamiento y test

Una vez que se dispone del corpus de críticas de cine elegido para la realización del proyecto (cuya estructura y contenido se explica detalladamente en el **apartado ¡Error! No se encuentra el origen de la referencia.**) el primer paso consiste en obtener os conjuntos de entrenamiento y test.

La tarea llevada a cabo para obtener los *set* de entrenamiento y evaluación para el diseño del clasificador se realiza a través de fichero *corpusCriticas.php*. En este fichero se han implementado una serie de funciones necesarias para el desarrollo de esta tarea, algunas de las cuales se detallan a continuación. Se debe tener en cuenta que las funciones descritas serán para el caso más general en el que se trabaja con 5 clases (Excelentes, Buenas, Indiferentes, Malas y Pésimas). Para el resto de casos considerados durante el proyecto, es decir el empleo de 4 clases (Excelentes, Buenas, Malas y Pésimas) y 2 clases (Positivas y Negativas), estas funciones serán modificadas adecuadamente aunque manteniendo básicamente su misma estructura y funcionalidad.

- ***obtenerFicherosCorpus()***: se almacena en un array *($ficheros_corpus)* los nombres de los ficheros disponibles.

- ***dividirFicheros():*** dentro de la variable *$ficheros_corpus* se localizan los ficheros que pertenecen a cada una de las categorías empleadas. Este proceso es sencillo ya que la primera letra del nombre de cada fichero se corresponde con la inicial de la clase a la que pertenece. Se ha optado por guardar el 70% de cada una de las categorías (todas disponen del mismo número de ficheros) en un directorio destinado a la fase de entrenamiento y el 30% restante de los ficheros en otro directorio que se empleará para la evaluación del sistema.

4.2.4.2 Preprocesado

En la fase de Preprocesado, como ya se explicó con anterioridad, se quieren obtener aquellos términos relevantes para el clasificador, de entre todos los que conforman cada una de los documentos.

Para obtener los resultados deseados tras la ejecución de esta etapa, en primer lugar se deben obtener los ficheros con el análisis morfológico de cada uno de las críticas (haciendo uso de la herramienta *STILUS Core* explicada en el **apartado 4.2.1.1**). Se ha decidido realizar esta tarea de forma aislada haciendo uso del fichero *analisisCriticas.php* puesto que solo es necesario realizarla una única vez para dejar almacenados en unos nuevos directorios los análisis de los textos empleados para el entrenamiento y para la evaluación del sistema.

- ***obtenerFicheros()***: Se obtienen los ficheros de los conjuntos de entrenamiento y test que se quieren emplear y se almacenan en las variables *$ficheros_train* y *$ficheros_test.*

- ***analisisFicheros():*** Cada fichero almacenado en las dos variables anteriores se procesa con ayuda de la herramienta *STILUS Core,* obteniéndose su análisis morfológico gracias a la ejecución del siguiente comando:

  > cat $fichero | /home/miracle/tools/stilus-core-es/stilus-core-es –d /home/miracle/tools/stilus-core-es/ -Mam –g >ficherosalida.txt

  Como se puede observar, la salida del comando se vuelca en un nuevo fichero para su posterior procesamiento. Los nombres de estos ficheros de salida, en función de la categoría a la que pertenezca el original, comenzarán por la inicial que identifica a dicha clase. De esta forma, por ejemplo, el análisis morfológico del primer fichero analizado de la clase *excelente* se almacenará como *e_análisis1.txt.*

Una vez que se tienen los análisis de los ficheros de entrenamiento y de test almacenados en dos nuevos directorios se puede pasar a seleccionar aquellas palabras que sean representativas para llevar a cabo el diseño del clasificador. Se abre y se lee en su totalidad de uno en uno todos los ficheros almacenados (bien sean de entrenamiento o de test) aplicando diversas funciones para obtener únicamente las palabras deseadas:

- ***obtenerAnalisisTrain()***: Se almacena en un array (*$ficheros_train*) los ficheros que conforman el conjunto de entrenamiento(los ficheros ya contiene el análisis morfológico de los mismos).

- ***obtenerAnalisisTest()***: Se almacena en un array (*$ficheros_test*) los ficheros a clasificar (los ficheros ya contiene el análisis morfológico de los mismos).

- ***obtenerFicheros_aux()***: Se almacena en sendos arrays, con el fin de poder trabajar con ellos, el contenido de los ficheros que contienen los listados de sinónimos y modificadores de la intensidad de las palabras.
- ***buscarModificadores():*** En caso de tratarse de una asignación de pesos basada en TF, se comprueba si una palabra se corresponde con uno de los modificadores de intensidad definidos (**Anexo2**) y en tal caso, se almacena el valor asociado a ese modificador y se activa un flag que indicará que se ha encontrado.

- ***comprobarCategoria():*** Se comprueba si el término del documento pertenece a una de la categorías gramaticales de interés (en este caso adjetivos, verbos y sustantivos) y de ser así se selecciona y se almacena dicha palabra. Para saber si la palabra es de interés bastará comprobar en el fichero de análisis si la etiqueta que corresponde a la categoría de la palabra es 'A', 'V' o 'N'.

- ***buscarSinonimo():*** Se comprueba si la palabra almacenada (de una de las categorías gramaticales de interés) se encuentra definida dentro del fichero de sinónimos construido *(Sinonimos.txt)*. En caso de encontrarse, se sustituye dicha palabra por el sinónimo que le corresponda (el primer elemento de la fila del conjunto de sinónimos en el que se encuentra dentro del fichero).

4.2.4.3 Entrenamiento

El resultado de esta etapa será una lista de palabras y una matriz de entrenamiento que serán empleadas durante la clasificación de los documentos de test. La lista de palabras contendrá aquellas presentes en uno o varios documentos del *set* de entrenamiento así como el número de textos diferentes en los que aparece. Por otro lado, la matriz de entrenamiento resultante contendrá el peso (binario o asociado a TF) correspondiente a cada palabra de la lista dentro de cada uno de los documentos del conjunto de entrenamiento.

Algunas de las funciones más importantes desarrolladas para implementar esta fase se detallan a continuación:

- ***añadirPalabras()***: Añade la palabra, seleccionada tras el preprocesado de uno de los documentos de entrenamiento, al vector provisional que representa a ese documento. Comprueba si dicha palabra existe ya en la lista general de palabras que se está formando dinámicamente (*$lista*). En caso de no encontrarse en dicha lista la almacena y en caso de que ya estuviera almacenada, se aumenta el contador asociado a la aparición de la palabra en los distintos documentos de entrenamiento.

- ***eliminarPalabrasLista()***: Se eliminan de la lista general todas aquellas palabras que sólo aparecen en un reducido número de documentos (porque serán muy específicas del mismo y no servirán para compararlos con otros) al igual que aquellas que aparezcan en demasiados textos de entrenamiento distintos ya que no servirán como diferenciación. Los umbrales elegidos para realizar la eliminación de las palabras pueden variarse con facilidad y serán un parámetro a estudiar durante la evaluación del sistema (apartado 5.3).

- ***añadirVectorMod()***: En caso de tratarse de clasificación basada en TF (no en el caso binario), se añade al vector de modificadores asociado a cada documento una entrada que corresponda al nombre (vector asociativo) de la palabra cuyo valor se encuentra incrementado o disminuido por la presencia de un modificador (se sabe si una palabra tiene la influencia de un modificador porque durante el preprocesado quedó registrado este hecho, gracias a la acción *buscar modificadores*). La posición asociada a la palabra modificada contendrá el valor previamente almacenado de dicho modificador (obtenido del fichero). Este vector de modificadores servirá posteriormente para incrementar o disminuir (durante la asignación de pesos) la importancia de la presencia de las palabras en el documento.

- ***añadirVectorEntrenamiento()***: Se recorre la lista general de palabras buscando cada una de ellas dentro de los vectores provisionales de los documentos de entrenamiento. Se añadirá una nueva fila (por cada documento) a la matriz de entrenamiento general que contendrá la presencia (caso binario) o la frecuencia (asignación TF) de las palabras de la lista general en el documento. En caso de tratarse de asignación basada en TF, al valor de la frecuencia de aparición de la

palabra en el documento se le añadirá (sumará o restará) el valor de los modificadores de la misma, si los tuviera.

4.2.4.4 Clasificación

Una vez realizada la fase anterior se dispone de la matriz de entrenamiento del sistema y de la lista de palabras relevantes a considerar.

Para obtener las clases de los documentos que servirán para la evaluación del sistema bastará con realizar la asignación de pesos de las palabras de la lista en cada uno de los documentos del conjunto de test y, posteriormente, calcular la distancia de cada vector de test con cada uno de los vectores que conforman la matriz de entrenamiento del sistema. La clase asignada será la que más aparezca entre los k documentos de entrenamiento cuyos vectores representativos sean más parecidos al de test. En caso de empate entre clases la asignada será aquella con el grado de positividad más alto dentro de la escala que se esté considerando.

Las funciones implementadas más destacadas de esta etapa se describen a continuación:

- ***crearVectorDoc()***: Al igual que en la fase anterior, se obtiene un vector provisional de cada documento del conjunto de test que contiene sólo aquellas palabras que han sido determinadas como relevantes tras el preprocesado.

- ***añadirVectorMod()***: Se trata de la misma función empleada durante el entrenamiento y que por tanto realiza las mismas funciones ya comentadas.

- ***obtenerVectorTest()***: Se construye el vector final representativo de cada documento de test que contendrá, en cada posición, la frecuencia de aparición (o la presencia para el caso binario) unida al efecto de los modificadores de cada una de las palabras de la lista general en el documento de test que se quiere clasificar.

- ***calcularDistancias()***: Se obtiene un vector que contiene la distancia (medida de similitud) entre cada documento de test y cada uno de los vectores que componen la matriz de entrenamiento del sistema.

- ***kNN():*** Se obtiene la clase de los k documentos de entrenamiento cuya distancia (almacenada en el vector de distancias salida de la función anterior) es menor con respecto al vector de test. La obtención de estas clases es sencilla puesto que los nombres almacenados de los vectores de entrenamiento comienzan siempre por la inicial de la clase a la que pertenecen. La clase asignada al documento de test que se desea clasificar será aquella a la que más documentos de entrenamiento pertenezcan de entre los k más parecidos seleccionados con anterioridad.

Una vez concluida la implementación de todo el sistema se puede proceder a su evaluación. Las pruebas y los resultados obtenidos se muestran en el **apartado 5.3**.

4.3. SISTEMA BASADO EN EL EMPLEO DE UN DICCIONARIO AFECTIVO

El sistema propuesto en esta ocasión está basado en la utilización de un diccionario afectivo, una de las técnicas (comentadas en el **apartado 2.2.3**) empleada con frecuencia en el campo de la minería de opinión. Este tipo de técnicas se basan en buscar las palabras afectivas que contiene un texto en un diccionario de vocablos afectivos construido previamente. La emoción global del texto se determina a partir de la media de los valores emocionales de cada una de las palabras clave detectadas.

Para la implementación de este sistema se ha optado por emplear un diccionario etiquetado semánticamente, el "*General Inquirer*[1]" del cual, a pesar de proporcionar palabras clasificadas dentro de otras muchas categorías, se han extraído sólo aquellas clasificadas dentro de una o varias de las categorías de interés: Positivas (*Positiv),* Negativas *(Negativ),* Fuertes (*Strong*) *o* Débiles *(Weak).* Para trabajar con el diccionario, éste estará almacenado en un fichero que contendrá aquellas palabras que resultan de interés seguidas, cada una de ellas, de un carácter que indicará la o las categorías a las que pertenece. Un fragmento del fichero que contiene el diccionario se muestra a continuación:

flojo N;pretencioso N;homofóbico N+;abusivo N;escatológico N;agotador N+;interesante P;lejos -;original P;llamativo +;ramplonería N+;marchito N;desangelado N;vergonzosa N;zafia +;soso N;artesano P;profesional P+;desamparar N-;abandonar N-;abyecto N;capaz P+;anodino N;anormal N;abominable N+;abundar P;abrasivo N+;raspante N+;brusco N;ausencia N-;falta N-;falto N-;faltar N-;absoluta +;absoluto +;absurdo N;absurdidad N;esperpento N;ridiculez N;abundancia P+;maremagno P+;magno P+;maremágnum P+;abundante P+;copioso P+;abusar N+;abuso N+;

Figura 15. Fragmento del diccionario afectivo empleado

[1] *http://www.wjh.harvard.edu/~inquirer/homecat.htm*

De esta forma, aquellas palabras presentes en el diccionario seguidas de "P" serán aquellas clasificadas como positivas y las que están seguidas de "N" aquellas consideradas con connotaciones negativas. Del mismo modo aquellas palabras que en su valoración presenten "+" serán consideradas como fuertes y aquellas seguidas de "-" como débiles. Una misma palabra solo puede ser considerada positiva o negativa por un lado y fuerte o débil por otro.

De acuerdo a lo explicado en el capítulo 3 y siguiendo los mismos pasos que en el diseño del sistema anterior (basado en el algoritmo kNN) se puede dividir el desarrollo de este nuevo sistema en tres etapas básicas: preprocesado, entrenamiento y clasificación.

Tanto los documentos pertenecientes al conjunto de entrenamiento como los pertenecientes al de test son introducidos en una fase de preprocesado, con el objetivo de obtener una representación estructurada de los mismos y seleccionar sólo aquellas palabras de los textos que sean de interés para la construcción del clasificador, es decir aquellas que se encuentren en el diccionario afectivo.

Como resultado de la fase de entrenamiento se obtendrán el o los umbrales que determinarán las clases a las que pertenecerán los documentos del conjunto de test. Estos últimos, en la etapa de clasificación, tras un preprocesado y una búsqueda de las palabras de interés de los mismos en el diccionario, presentarán un valor global de carga semántica del documento, de modo que en función de los umbrales de decisión (fruto del entrenamiento) serán asignados a una u otra categoría de entre todas las posibles.

Se debe tener en cuenta que este caso el corpus empleado será básicamente el mismo que el utilizado en el sistema anterior con la particularidad de que estará adaptado (ver **apartado 5.1**) para contemplar solo dos opciones: una clasificación en dos clases (positiva y negativa) o una clasificación en cuatro clases (excelente, buena, mala y pésima). En el desarrollo de este sistema no se contempla el uso de las críticas indiferentes ya que estas presentan un vocabulario muy específico de la clase, no son algo sencillamente neutral o intermedio ente las críticas positivas y negativas lo cual provoca que únicamente haciendo uso del diccionario no se pueda establecer un umbral adecuado para diferenciar documentos de esta categoría.

4.3.1. PREPROCESADO

Este módulo es ejecutado tanto en la fase de entrenamiento como en la de test y su objetivo no es otro que el de obtener una representación estructurada y simplificada de los documentos que ayuden a su posterior procesado. Tras la ejecución de esta etapa se debe obtener de cada texto todas aquellas palabras representativas del mismo y que, por tanto, sean de interés para llevar a cabo la clasificación.

4.3.1.1 Análisis sintáctico y lematización

La necesidad de esta etapa ya ha sido comentada en el diseño del sistema anterior (el basado en kNN) y su implementación se ha llevado a cabo siguiendo prácticamente los mismos pasos ya explicados para ese sistema (**apartado 4.2.1.1**).

Para llevar a cabo un análisis morfológico de los documentos, que permita extraer solo aquellos términos que de verdad resulten relevantes para la representación de los mismos, se ha empleado (al igual que en diseño del sistema kNN) la herramienta de procesado lingüístico ***STILUS Core*** (cuyo funcionamiento se explicó con detalle en el preprocesado del sistema anterior). Como resultado de este análisis se obtienen los ficheros asociados a cada documento que contienen un análisis morfológico del que se puede obtener la raíz de cada una de las palabras que conforman el documento y la categoría gramatical de éstas (previo proceso de desambiguación realizado).

4.3.1.2 Obtención de sinónimos

Como ya se comentó para el diseño del sistema anterior (**apartado 4.2.1.2**), además de poder seleccionar sólo aquellas palabras que pertenecen a las categorías gramaticales de interés, y que por tanto son útiles para la clasificación del documento, se ha elaborado un archivo auxiliar (**Anexo 1**) que contiene sinónimos de muchas de las palabras con una posible carga afectiva que, presumiblemente, pueden aparecer en las críticas analizadas.

De esta forma, tras el preprocesado de los documentos se dispone del lema de todas las palabras que lo conforman (o su sinónimo representativo en caso de tenerlo) las cuales, a priori, son relevantes o útiles para el proceso de clasificación de los documentos.

4.3.2. ENTRENAMIENTO

Una vez realizado el preprocesado de los documentos del conjunto de entrenamiento, se pasa a entrenar el sistema con el fin de obtener el o los **umbrales** que determinarán la categoría a la que se asignarán los documentos de test que se emplearán para la evaluación.

Para cada uno de los documentos que forma parte del *set* de entrenamiento se realizan las mismas acciones a fin de obtener el valor "afectivo" asociado a cada uno de ellos.

4.3.2.1 Búsqueda en el diccionario y valoración

En primer lugar, se buscan todas las palabras representativas del documento (obtenidas tras el preprocesado) en el **diccionario afectivo** del que se dispone. Si la palabra existe en dicho diccionario, entonces se obtiene su valoración (P, N, +, -) la cual se encuentra inmediatamente después de la misma. Se ha definido que por cada palabra cuya valoración sea "*P*" o "*N*" se sume o se reste uno respectivamente al valor afectivo global del documento. Del mismo modo, por cada valoración correspondiente a "*+*" o "*-*" se sumará o restará uno respectivamente al indicador de intensidad afectiva del mismo. Se actualiza el valor del documento a medida que se procesan las palabras que lo conforman de manera que, al final, se obtiene un resultado global de la orientación (positividad o negatividad del documento) y la intensidad (o fuerza) del texto procesado.

4.3.2.2 Cálculo de umbrales de decisión

Cada vez que se obtiene la valoración global (de orientación e intensidad) de uno de los documentos del conjunto de entrenamiento, este valor se añadirá al valor almacenado para la categoría a la que dicho documento pertenezca. Gracias a esto, una vez procesadas todas las críticas, se puede realizar una media de la valoración (tanto de orientación como de intensidad) de todas las categorías con las que trabaja el sistema.

Estas medias serán los umbrales de decisión que servirán para la implementación del clasificador y por tanto para la determinación de las clases a las que pertenecen cada uno de los documentos del conjunto de test.

De esta forma la media de las valoraciones afectivas globales de cada una de las clases (tanto en orientación como en intensidad) conformarán los umbrales de decisión del sistema. En la **Figura 16** puede observarse que en el caso de la clasificación en dos categorías sólo existe un único umbral (U_{PosNeg}) que será calculado teniendo en cuenta únicamente los valores globales de orientación de las clases de entrenamiento. En el caso de existir cuatro posibles categorías el valor del umbral (U_{PosNeg}) que distingue entre documentos positivos (excelentes y buenos) y negativos (malos y pésimos) se calcula de la misma forma que en el caso anterior, basándose en el valor global de orientación de las distintas clases. Por otro lado, en este caso, para determinar la diferencia entre los documentos que serán clasificados como buenos o excelentes (U_{be}) y como malos o pésimos (U_{mp}) se empleará la media de los valores globales de intensidad afectiva de los documentos de entrenamiento.

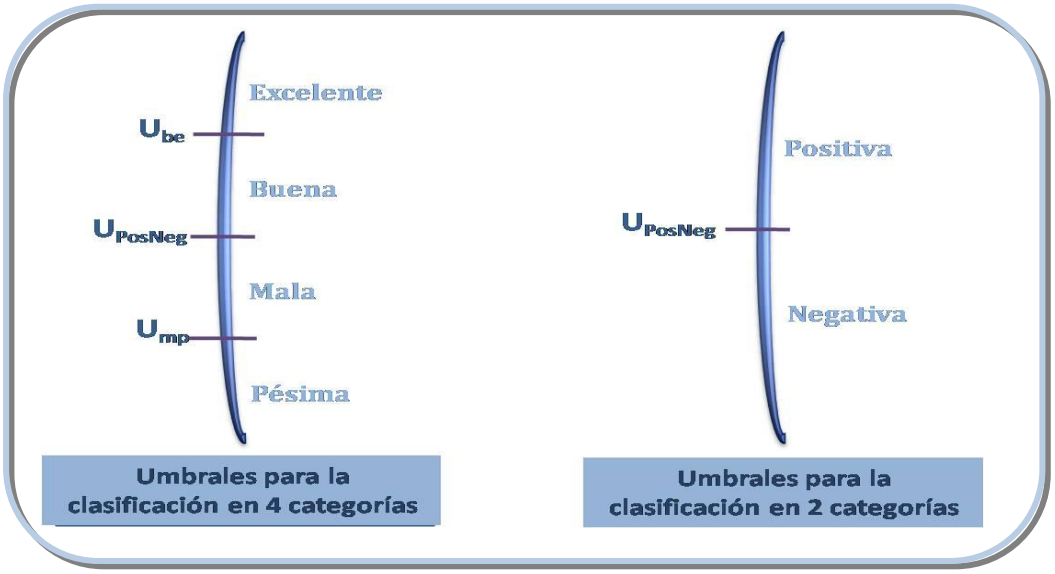

Figura 16. Umbrales de decisión del clasificador

4.3.3. CLASIFICACIÓN

En este modulo se aborda el problema de clasificar cada uno de los documentos que componen el conjunto de test dentro de una de las categorías definidas. Una vez obtenidos los umbrales de decisión, gracias al análisis de los documentos de entrenamiento, se puede hacer uso de estos umbrales para llevar a cabo la clasificación de nuevos elementos. Así, la resolución de la consulta consiste en establecer o determinar el valor (en términos de orientación e intensidad) de cada uno de los documentos de test para, de esa forma, poder determinar la categoría a la que pertenecen.

4.3.3.1 Búsqueda en el diccionario y valoración

Del mismo modo en que se ha procedido con los documentos del *set* de entrenamiento, se realiza para cada una de las palabras representativas del texto a clasificar, obtenidas tras el preprocesado, una búsqueda de las mismas en el diccionario afectivo del que se dispone. En función de la valoración que tengan las palabras (en caso de existir en el diccionario) se calcula sucesivamente el valor afectivo del documento tanto en términos de orientación (número de *"P" o "N"*) como en términos de intensidad (número de "+" o "-"). Una vez analizadas todas las palabras que componen uno de los documentos de test se obtiene una valoración global del texto que permitirá clasificarlo adecuadamente.

4.3.3.2 Determinación de categorías

El proceso para asignar una categoría a los documentos que componen el conjunto de test es muy sencillo, gracias a los umbrales de decisión obtenidos tras el entrenamiento del sistema.

Puesto que se dispone de la valoración global del documento, este valor se compara con el o los umbrales definidos para el sistema y se asigna el texto a la categoría definida para ese valor con respecto a los umbrales de decisión.

Se han diseñado distintas opciones para la utilización del sistema (haciendo uso de la orientación y la intensidad de los documentos) en función del número de categorías que deseen contemplarse, es decir en caso de querer sólo diferenciar entre

críticas positivas o negativas (2 clases) o si se desea diferenciar entre orientaciones más concretas dentro de la positividad y negatividad de los textos.

Todos los casos contemplados pueden observarse en el apartado que corresponde a la validación del sistema (**apartado 5.3**) pero, a grandes rasgos, se puede decir que en caso de tratarse de una clasificación binaria o clasificación por polaridad de la opinión (opiniones positivas o negativas) solo se emplea el valor de la orientación (número de "P" y "N") de cada documento y, sin embargo, para el caso de una clasificación en la que debe diferenciarse el grado de positividad o negatividad (varias clases) se hace uso también de la valoración del documento en términos de la intensidad del mismo (presencia de de los parámetros "+" y "-").

Para ver gráficamente estos casos explicados podemos considerar un ejemplo (**Figura 17**) en el cual un texto del conjunto de test cuya clase real es *excelente* es clasificado correctamente por el sistema (como excelente o positivo) en función de que se use la clasificación en dos clases o en cuatro posibles clases.

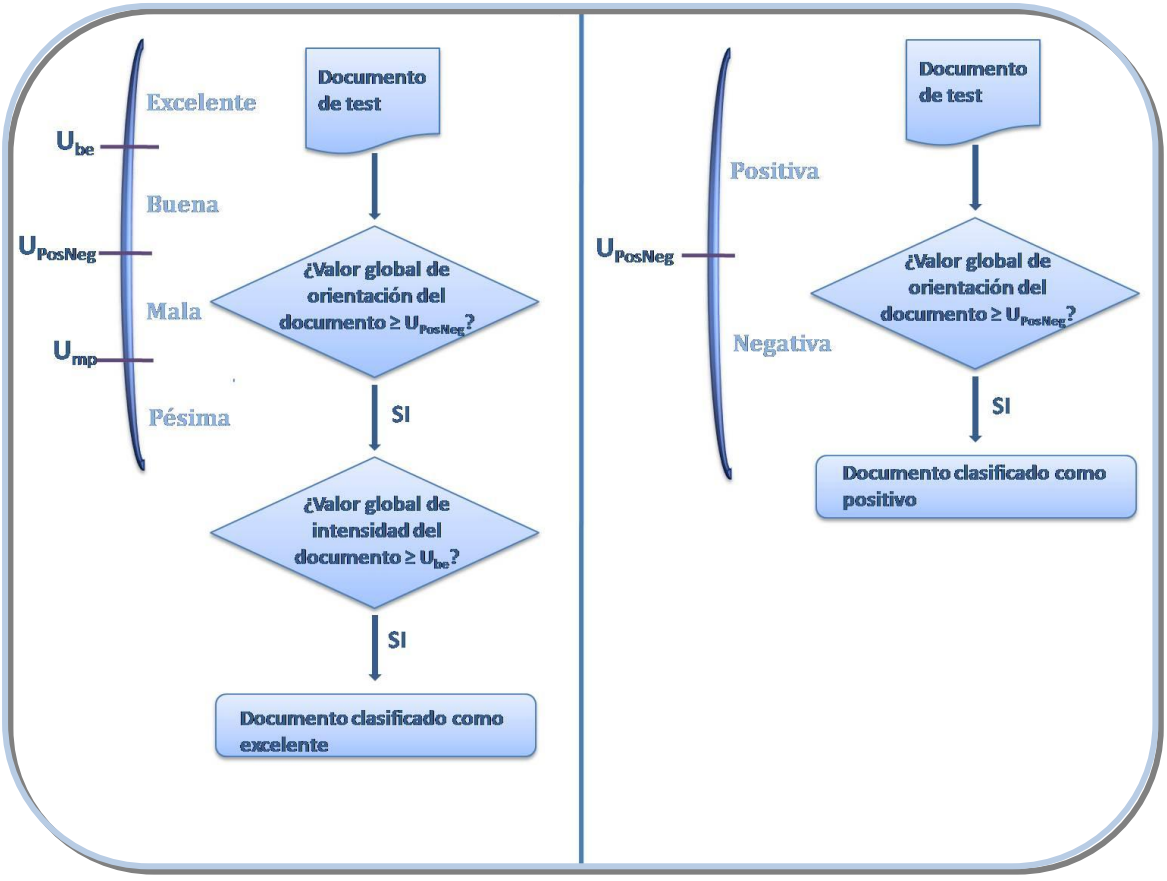

Figura 17. Ejemplo de clasificación basada en diccionario afectivo

4.3.4. IMPLEMENTACIÓN

El sistema de clasificación descrito, al igual que el sistema anterior, ha sido implementado en el lenguaje de programación **PHP** (*PHP Hypertext Pre-processor*), por su sencillez, versatilidad y potencia. En las siguientes secciones se explicarán las funciones implementadas para cada uno de los módulos y, por tanto, aquellas necesarias para poder llevar a cabo el diseño del clasificador.

El diagrama general del sistema se puede observar en la siguiente figura.

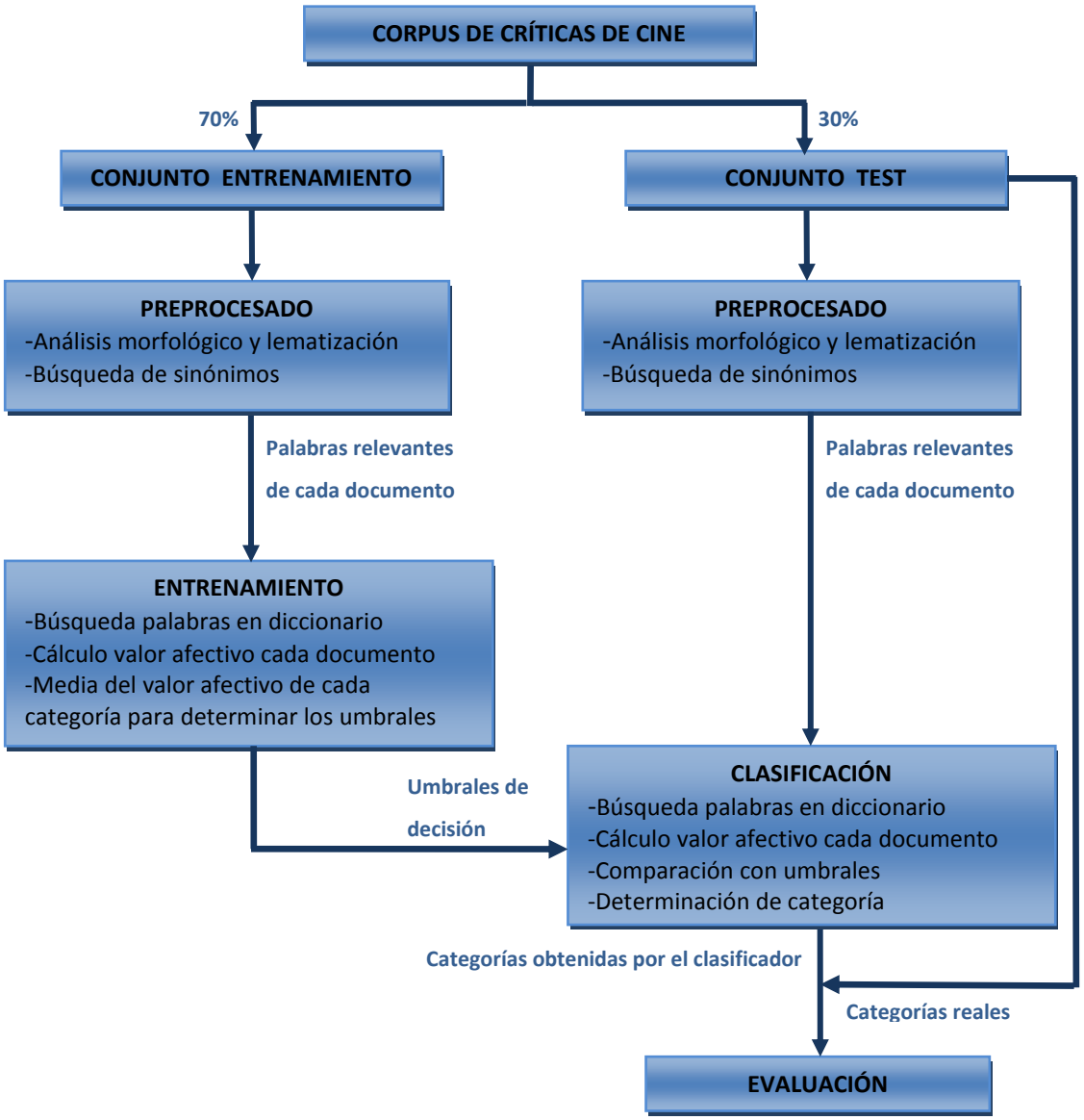

Figura 18. Diagrama sistema basado en diccionario afectivo

4.3.4.1 Conjunto de entrenamiento y test

Una vez que se dispone del corpus de críticas de cine elegido para la implementación del sistema (adaptado, como ya se ha comentado, para no tener en cuenta las críticas indiferentes), el primer paso consiste en obtener el conjunto de entrenamiento y el de test para la implementación del sistema.

La tarea llevada a cabo para obtener el *set* de entrenamiento y test así como las funciones necesarias para su implementación se explicaron en la implementación del sistema anterior (ver **apartado 4.2.4.1** para más detalle) y es llevada a cabo a través del fichero *corpusCriticas.php*. A modo de resumen, se puede decir esta tarea consiste en almacenar en un directorio el 70% de los documentos de cada una de las clases con las que queremos trabajar (puede ser una clasificación binaria o multiclase) para conformar el conjunto de entrenamiento y el 30% restante en otro directorio para constituir el *set* de test que servirá para realizar la evaluación del sistema.

4.3.4.2 Preprocesado

En esta fase, como ya se explicó con anterioridad, se quieren obtener aquellos términos relevantes para el clasificador, de entre todos los que conforman cada una de los documentos.

En primer lugar se deben obtener los ficheros que contengan el análisis morfológico de cada uno de las críticas para lo cual se hace uso de la herramienta *STILUS core* (explicada con detalle en el **apartado 4.2.4.1**). Esta tarea, como se explicó en la implementación del sistema basado en kNN, se ha realizado de forma aislada haciendo uso del fichero *analisisCriticas.php* puesto que solo es necesario realizarla una única vez para dejar almacenados en unos nuevos directorios los análisis de los textos empleados para el entrenamiento y para la evaluación del sistema. Las funciones implementadas para la realización de esta tarea se explicaron con detalle en la descripción de la implementación del sistema anterior (ver **apartado 4.2.4.2**).

Una vez que se tienen los análisis de los ficheros de entrenamiento y los de test almacenados en dos nuevos directorios se puede pasar a seleccionar aquellas palabras que sean representativas para llevar a cabo el diseño del clasificador. Se abre y se lee en su totalidad de uno en uno todos los ficheros disponibles (bien de entrenamiento o bien de test) aplicando diversas funciones para obtener únicamente las palabras deseadas:

- ***obtenerDiccionario()***: Se almacena en un vector (*$diccionario*) el contenido del fichero que contiene el diccionario del que se dispone para implementar el clasificador. De esta forma, en el vector está almacenado en una posición una palabra y en la posición inmediatamente posterior la valoración que le corresponde.

- ***obtenerFicherosTrain()***: Se almacena en un array (*$ficheros_train*) el nombre de los ficheros del conjunto de entrenamiento (los ficheros ya contiene el análisis morfológico de los mismos).

- ***obtenerFicherosTest()***: Se almacena en un array (*$ficheros_test*) el nombre de los ficheros a clasificar (los ficheros ya contiene el análisis morfológico de los mismos).

- ***obtenerFichero_Sinonimo()***: Se almacena en un array, con el fin de poder trabajar con él, el contenido del fichero donde se encuentra definido el listado de los sinónimos determinados para el sistema.

- ***comprobarSinonimo()***: Se busca la palabra que corresponda, del fichero devuelto por el *STILUS Core* que se está procesando, en el listado de sinónimos existente y, en caso de encontrarse, se sustituye la misma por el sinónimo definido como representativo para esa palabra.

De esta forma, tras el preprocesado de los documentos se dispone del lema de todas las palabras que lo conforman (o su sinónimo representativo en caso de tenerlo) las cuales, a priori, son relevantes o útiles para el proceso de clasificación de los documentos.

4.3.4.3 Entrenamiento

El resultado de esta etapa es el o los umbrales que serán empleadas durante la clasificación de los documentos de test para definir la categoría a la que pertenece cada uno de ellos. Los umbrales serán la media de la valoración de los documentos de entrenamiento para cada una de las categorías definidas. Algunas de las funciones más importantes desarrolladas para implementar esta fase se detallan a continuación:

- ***buscarDiccionario()***: Se busca cada una de las palabras representativas del documento (devueltas tras el preprocesado) en el diccionario del que se dispone. En caso de encontrarse, se devuelve la valoración (P, N,+,-) que corresponde a dicha palabra en el mismo (tal valoración se encuentra almacenada inmediatamente después de la palabra en el vector del diccionario, $Diccionario).

- ***actualizarValorFich()***: A medida que se obtiene la valoración de cada una de las palabras representativas de un documento, se actualiza el valor global del mismo. De esta forma, se sumará o se restará uno al valor global de la orientación del fichero (en función de si se encuentra "*P*" o "*N*" respectivamente) y se sumará o se restará uno a la intensidad global del fichero (en función de si en el diccionario hay almacenado "+" o" –" asociado a la palabra que corresponda).

- ***actualizarCategorias()***: Una vez procesadas todas las palabras representativas de uno de los documentos de entrenamiento, se actualiza el valor global de la categoría a la que pertenezca dicho documento, sin más que añadir al contador de la categoría en cuestión el valor global obtenido para ese documento.

- ***obtenerUmbral()***: A partir de los valores almacenados asociados a cada categoría, una vez analizados todos los documentos de entrenamiento, se obtiene la media de cada una de las clases, tanto en orientación como en intensidad. Dichas medias son los umbrales de decisión que servirán para clasificar adecuadamente nuevos documentos. Estos umbrales serán, por tanto, la salida de este módulo de entrenamiento y la entrada a la fase de clasificación.

4.3.4.4 Clasificación

Una vez realizada la fase anterior se dispone de un vector que almacena el o los umbrales de decisión (según el tipo de clasificación del que se trate) que determinarán a qué categoría pertenecen nuevos documentos de entrada al sistema. Del mismo modo se obtiene tras una fase de preprocesado de cada uno de los documentos aquellas palabras (o sus sinónimos) que son representativas del mismo.

Para obtener la clase de los nuevos documentos, que servirán para la evaluación del sistema, se debe obtener la valoración global de cada uno de ellos, de acuerdo a la valoración que en el diccionario se da para las palabras que componen el texto. Una vez obtenida dicha valoración, se compara con los umbrales de decisión obtenidos tras el entrenamiento y se determina la clase del nuevo documento en función de su valor con respecto a los umbrales de decisión.

Algunas de las funciones más importantes desarrolladas para implementar esta fase se detallan a continuación (algunas de las funciones para encontrar la valoración global de los documentos de test serán las mismas empleadas en el tratamiento de los textos del conjunto de entrenamiento):

- ***buscarDiccionario()***: Se trata de la misma función explicada en el apartado anterior y empleada con los documentos de entrenamiento.
- ***actualizarValorFich()***: Se actualiza el valor afectivo global del fichero de la misma forma que se explicó para los documentos de entrenamiento.
- ***obtenerCategorías()***: A partir del vector devuelto tras la obtener la valoración de cada uno de los ficheros de test, se calcula y se devuelve la categoría a la que pertenece cada uno de ellos, comparando dichos valores con el valor del o de los umbrales obtenidos tras la fase de entrenamiento. Según el modelo del clasificador con el que se está tratando, dos clases (positiva y negativa) o cuatro clases (excelente, buena , mala y pésima), se emplea para dicha comparación con los umbrales únicamente el valor del fichero que corresponde a su orientación afectiva (grado de positividad) o se emplea también el valor que corresponde a la intensidad de la opinión

5. VALIDACIÓN DEL SISTEMA

5.1. DESCRIPCIÓN DEL CORPUS

Para poder llevar a cabo el desarrollo del proyecto primero es necesario contar con el recurso adecuado, en este caso contar con un corpus de críticas de cine en español. Puesto que no se tiene constancia de la existencia de ningún recurso de la naturaleza requerida para el español y, aunque así fuera, con el fin de poder asegurar la fiabilidad del mismo, se optó por la **generación de un corpus propio**.

Para acometer la construcción del corpus empleado en el proyecto era necesario encontrar alguna web que se dedicara a este tema de la cual poder extraer el corpus de forma automática. Las características que, a priori, se decidió que debía cumplir dicha web eran:

- Un número elevado de críticas disponibles de distintas categorías.
- En el caso de ser contenidos generados por los usuarios, que se asegurara una mínima calidad de los textos, en cuanto a lo que se refiere a faltas de ortografía, vocabulario empleado y correcta escritura de las palabras empleadas.
- La presencia de una puntuación o valoración asociada a cada crítica asignada por el autor de la misma.

Las críticas de cine contenidas en la mayoría de las webs son introducidas en la misma por usuarios y no por críticos especializados. Esto añade un punto de dificultad a la tarea que se está realizando puesto que los textos pueden contener faltas de ortografía, incoherencias entre lo que se relata y la puntuación final asignada, divergencia entre los tamaños de las distintas críticas, etc.

Bajo las condiciones deseadas descritas anteriormente, aunque se encontraron más webs que también las cumplían, las elegidas para la construcción del corpus fueron *Muchocine*[2] y *Test&Vote*[3]. La elección de estas páginas se basó, básicamente, en la calidad y longitud de las críticas que contienen así como la clasificación de todas ellas en multitud de categorías que facilitan su búsqueda y elección (todas las críticas pertenecientes a una misma película aparecen juntas y, a su vez, cada película aparece clasificada, por ejemplo, por género y orden alfabético).

Una de las condiciones impuestas a la web de la que iban a ser extraídas las críticas era que éstas presentaran una valoración a modo de resumen (como puede ser un número de estrellas); esto, a priori, evitaría la tarea de etiquetar manualmente los documentos para el aprendizaje y bastaría con identificar cada una de las valoraciones posibles (por ejemplo de 1 a 5 estrellas) con cada una de las categorías que se desea obtener con el clasificador desarrollado. Pronto se pudo observar, al analizar algunas de las críticas encontradas, que existían muchas incoherencias entre la opinión expresada y la valoración resumen (o número de estrellas) asignada a dichas críticas; por ejemplo, en un elevado número de críticas siempre aparece una valoración media (3 estrellas) asociada a la opinión a pesar de tratarse de críticas con una clara orientación positiva o negativa. Se consideró, y se comprobó posteriormente, que esto podía ser un grave problema a la hora de implementar el clasificador con lo que, a pesar de tratarse de una tarea mucho más costosa, se optó por clasificar y etiquetar manualmente todas las críticas empleadas para el corpus. Este es uno de los motivos por el cual el corpus empleado finalmente no contiene un número de críticas muy elevado aunque sí suficiente como para desarrollar adecuadamente el entrenamiento del clasificador y su posterior evaluación.

[2] *http://www.muchocine.net/criticas.php*

[3] *http://www.testandvote.es. Esta web dejó de existir un tiempo después del comienzo del desarrollo del proyecto. No obstante, la elaboración del corpus ya se había completado y muchas de las críticas empleadas fueron obtenidas de esta fuente.*

Finalmente, el corpus completo empleado para la realización del proyecto (**Figura 19**) consta de un total de 500 críticas de cine, de las cuales 100 pertenecen a cada una de las categorías contempladas durante el desarrollo de los sistemas; tales categorías son: excelente, buena, indiferente, mala y pésima.

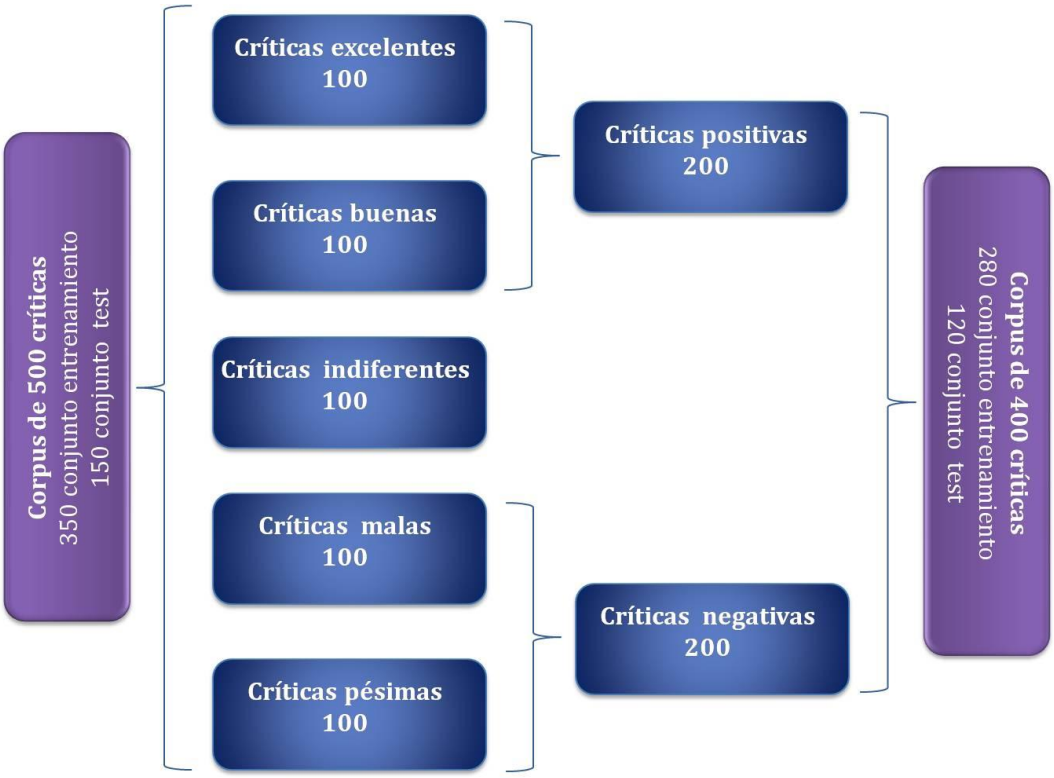

Figura 19. Estructura del corpus empleado

Como se verá más adelante, en los resultados del proyecto (**apartado 5.3**), se han contemplado diferentes opciones para la evaluación del clasificador para las cuales se adaptará adecuadamente el corpus disponible. Un primer caso consistirá en obtener un clasificador (implementado sólo haciendo uso del algoritmo kNN) capaz de diferenciar entre cinco categorías (excelente, buena, indiferente, mala y pésima); para su implementación y evaluación se hará uso del corpus completo disponible, es decir 500 críticas, de las cuales 100 pertenecen a cada una de las clases. Un segundo caso consistirá en evaluar un clasificador capaz de distinguir entre cuatro categorías (excelente, buena, mala y pésima) para lo cual se empleará el mismo corpus que en el caso anterior con la diferencia de que no se tendrán en cuenta las 100 críticas indiferentes (por tanto el corpus se compondrá de 400 documentos). Para finalizar, un último caso consistirá en la evaluación de un clasificador binario capaz de distinguir

únicamente entre opiniones positivas (aquellas del corpus etiquetadas como excelentes o buenas) y negativas (aquellas etiquetadas como malas o pésimas). En éste último caso el corpus considerado estará compuesto, como en el caso anterior, por 400 críticas de las cuales 200 serán positivas y 200 negativas.

Todas las críticas que componen el corpus se encuentran clasificadas dentro de una única categoría, de las cinco posibles que contempla el sistema en su caso más amplio; la forma de identificar la clase a la que pertenecen es muy sencilla ya que el nombre de los ficheros que las contienen comienza por la inicial correspondiente a la categoría que tienen asignada, así el primero fichero almacenado que contiene una opinión excelente se llamará *e_opinion1.txt*

El tamaño medio de las opiniones consideradas es de 60 palabras. Este es uno de los motivos por los que se eligieron las críticas de las webs comentadas, ya que en otros casos se trataba de textos de mucha mayor longitud que no eran propiamente de opinión si no que contenían otros elementos puramente objetivos, como pueden ser, por ejemplo, el argumento de la película en cuestión o la filmografía de los actores.

Para ilustrar el contenido de las críticas que componen el corpus, a continuación se presenta un ejemplo de cinco de ellas, cada una perteneciente a una categoría distinta de las contempladas a lo largo del proyecto.

La dirección es una auténtica lección de hacer cine, plano a plano, secuencia secuencia, sin perder en ningún momento el ritmo, sencillamente perfecto. En el film además contamos con una esplendorosa interpretación del genial Eastwood, con un personaje que tiene miles de matices en sus gestos austeros y en su mirada penetrante. Ejemplar es también la dirección de actores, que saca todo lo mejor de cada uno de los actores, sencillamente sobrecogedores ¿Se puede hacer mejor? NO. Es una película prodigiosa, modesta en su presupuesto, pero grande, monumental, inconmensurable en sus resultados. Una verdadera maravilla que pasará a la historia del cine.

Figura 20. Ejemplo de crítica *excelente*

Divertida, fresca y entretenida. Así se podría definir la película que nos ocupa. Numerosos gags divertidos nos llevan a lo largo de un metraje cargado de acción que logra tenernos en la butaca, rendidos a las aventuras de unos espías tan dispares. Recomendada para todos aquellos que quieran pasar un buen rato en estas fechas y que disfrutan de las comedias que no recurren necesariamente a chistes soeces para soltar la carcajada.

Figura 21. Ejemplo de crítica *buena*

Los pasajeros del tiempo, en definitiva, es una película que se toma el viaje en el tiempo con mucha frivolidad, que tiene una premisa muy imaginativa, pero que desaprovecha gran parte de su potencial y no cumple con las expectativas. Aunque si uno no es muy exigente, aun puede pasar una tarde de domingo bastante entretenida.

Figura 22. Ejemplo de crítica *indiferente*

Considero que el argumento es más adecuado para cómic donde quizás se podría desarrollar mucho más dándole el interés que en ningún momento el film consigue provocar. El guión, con unos diálogos previsibles y facilones y con un desenlace en cierto modo "soso", hace que si algo de este film tenía que pasar a la historia pase totalmente desapercibido al espectador.

Figura 23. Ejemplo de crítica *mala*

Esta es sin duda una de las peores películas que he visto en muchos años. El guión es lamentable, los diálogos patéticos y las actuaciones aun peores. Las dos chicas dan pena y el pobre Elijah sigue con la misma cara que llevaba cuando subia el monte del destino en Mordor. A los diez minutos de película el aburrimiento es tal que lo que deseas es que el próximo cadáver sea el del director. Haceos un favor y no perdáis el tiempo con este bodrio.Asi que no lo recomiento para nada.

Figura 24. Ejemplo de crítica *pésima*

5.2. MEDIDAS DE EVALUACIÓN

La evaluación del clasificador constituye la etapa final del sistema implementado. Dada una consulta, lo que se persigue es interrogar al conjunto de documentos con el fin de obtener una respuesta, en este caso, en forma de una categoría asignada a dicha consulta. Para saber en qué medida la respuesta obtenida es satisfactoria es necesario realizar una evaluación de los resultados. Cualquier sistema debe ser sometido a una evaluación para que los usuarios puedan comprobar cuál es su efectividad y pueda ser comparado de un modo fiable con otros sistemas similares.

Al igual que ocurre en los sistemas de RI, la evaluación de los clasificadores se realiza de forma experimental, en lugar de realizarse analíticamente; esto se debe a que en este último caso se necesitaría una especificación formal del problema a resolver. La evaluación experimental se encarga de medir la efectividad de los clasificadores, es decir su habilidad para tomar decisiones correctas.

Tradicionalmente, la efectividad de las operaciones en IR se mide utilizando las medidas clásicas de precisión (*precision*) y relevancia (*recall*). Esta práctica se ha seguido también en trabajos de categorización, aunque, en algunos casos, se ha optado por presentar los resultados en términos de porcentajes de aciertos y fallos.

La precisión con respecto a la categoría c_i ($\mathbf{p_i}$) se define como la probabilidad condicional $P(\breve{\phi}(d_x,c_i) = T|\phi(d_x,c_i) = T)$, es decir que si un documento d_x se asigna a la categoría c_i la predicción es correcta. De forma análoga, la relevancia con respecto a la categoría c_i ($\mathbf{r_i}$) se define como la probabilidad $P(\phi(d_x,c_i) = T|\ \breve{\phi}(d_x,c_i) = T)$, es decir, la probabilidad de que si un documento d_x debe ser clasificado bajo la categoría c_i esto suceda.

Estas probabilidades son estimadas en términos de la tabla de contingencia (**Tabla 3**) para la categoría c_i sobre el conjunto de test:

Tabla 3. Tabla de contingencia para una determinada categoría

Categoría c_i		Criterio del Experto	
		SI	NO
Criterio del Clasificador	SI	TP_I	FP_I
	NO	FN_I	TN_I

Donde TP_i (verdaderos positivos - *true positives*) representa el total de documentos que han sido correctamente clasificados en la categoría c_i; FP_i (falsos negativos – *false negatives*) es el número de documentos del conjunto de test que han sido clasificados dentro de c_i cuando su categoría es otra; FN_i (falsos negativos - *false negatives*) es el número de documentos pertenecientes a c_i que no han sido clasificados como tal; y TN_i (verdaderos negativos - *true negatives*), documentos que no perteneces a c_i y que no han sido asignados como tal.

La estimación de la precisión y la relevancia se obtiene de la siguiente manera:

$$\hat{p}_i = \frac{TP_i}{TP_i + FP_i}$$ **Ecuación 8**

$$\hat{r}_i = \frac{TP_i}{TP_i + FN_i}$$ **Ecuación 9**

Lo interesante es combinar ambas medidas en un único valor numérico utilizando la media armónica y eso es precisamente lo que hace la medida F_β:

$$F_\beta = \frac{(\beta^2 + 1) \cdot p \cdot r}{(\beta^2 \cdot p + r)}$$ **Ecuación 10**

β es un parámetro que permite estimar la influencia relativa de ambas medidas: precisión y cobertura. Si se considera proporcionar igual peso a ambas características (β = 1) la medida final a considerar que determinará las prestaciones del clasificador será la siguiente:

$$F = \frac{2 \cdot p \cdot r}{(p + r)}$$ **Ecuación 11**

Para obtener las estimaciones de la precisión y la cobertura se pueden emplear dos métodos:

- ***Micro-averaging***: la precisión (p) y la cobertura (r) se obtienen sumando todas las decisiones individuales.

$$p^{\mu} = \frac{TP}{TP+FP} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|}(TP_i+FP_i)}$$ **Ecuación 12**

$$r^{\mu} = \frac{TP}{TP+FN} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|}(TP_i+FN_i)}$$ **Ecuación 13**

- ***Macro-averaging***: la precisión (p) y la cobertura (r) se evalúan en primer lugar de forma local para cada categoría, y después se hace la media con los resultados para las diferentes categorías.

$$p^{M} = \frac{\sum_{i?1}^{|C|} p_i}{|C|}$$ **Ecuación 14**

$$r^{M} = \frac{\sum_{i?1}^{|C|} r_i}{|C|}$$ **Ecuación 15**

Mientras que el micro-averaging cada documento recibe igual peso (es una medida centrada en el documento), en el macro-averaging las categorías reciben igual peso (medida centrada en la categoría).

5.3. RESULTADOS DE LA EVALUACIÓN

Los resultados obtenidos tras la evaluación se presentarán divididos en dos grandes secciones, cada una de las cuales se corresponde con uno de los dos sistemas evaluados, el basado en el algoritmo kNN y el basado en el empleo del diccionario afectivo.

En la **Figura 25** se presentan, de manera general, las pruebas realizadas para la evaluación de los sistemas. En primer lugar, se evaluará el sistema basado en el algoritmo kNN, justificando todos los bloques diseñados para el caso de tener cinco categorías (excelente, buena, indiferente, mala y pésima). Se evaluará también este sistema si se tienen cuatro categorías posibles (sin críticas indiferentes) y si únicamente se consideran dos (críticas positivas o negativas). En todos los casos se considerará la asignación de pesos de forma binaria o empleando la frecuencia de aparición de los términos (asignación TF). En segundo lugar, se evaluará el sistema basado en el empleo del diccionario afectivo. En este caso, se obtendrán los umbrales necesarios para su implementación y se presentarán los resultados considerando el caso de tener dos categorías (positiva y negativa) o cuatro categorías (excelente, buena, mala, pésima).

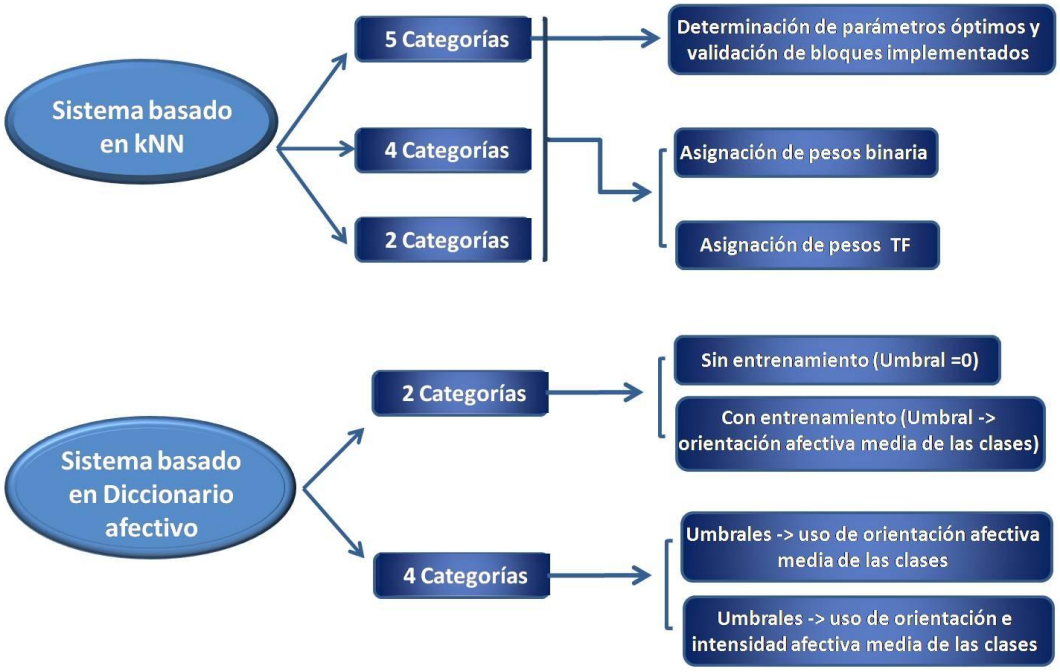

Figura 25. Esquema de pruebas realizadas

5.3.1. SISTEMA BASADO EN EL ALGORITMO KNN

En primer lugar se van a presentar los resultados obtenidos para el sistema implementado basado en el algoritmo kNN y, como punto de partida, se van a presentar las pruebas y los resultados que justifican la elección del valor de los distintos parámetros variables del sistema (valor de k y valor del umbral para eliminar palabras cuya frecuencia de aparición en documentos diferentes es muy baja).

Para decidir el valor óptimo del número de documentos de entrenamiento (k) que deben emplearse para determinar la categoría de las críticas del set de evaluación, se obtienen las prestaciones del sistema, empleando macro-averaging, (precisión, cobertura y medida F) a medida que varía el número de vecinos considerados para realizar la clasificación.

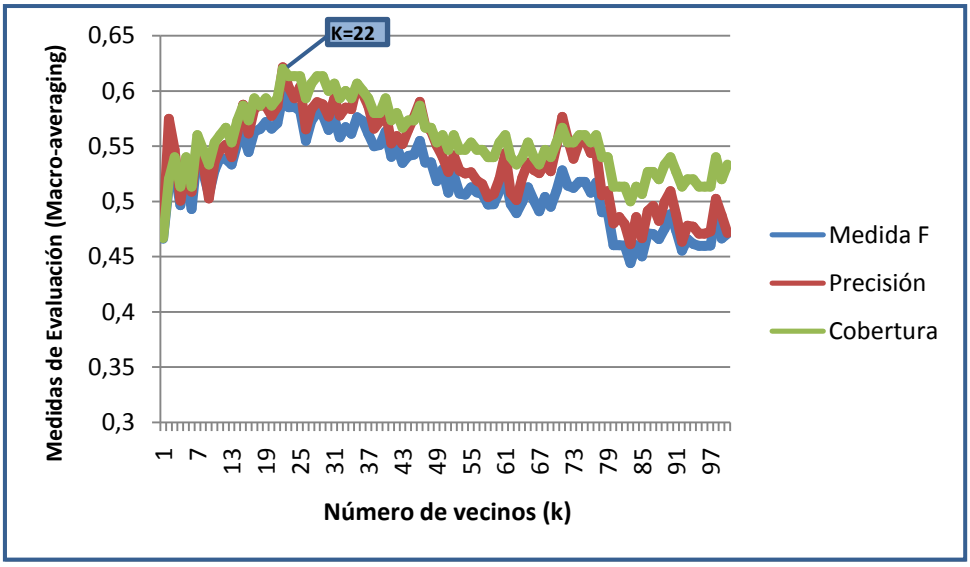

Figura 26. Evolución de las prestaciones en función de k

En la figura anterior se puede observar la evolución de las prestaciones del sistema a medida que aumenta el valor de la k para el caso de considerar 5 categorías posibles en la clasificación (excelente, buena, indiferente, mala y pésima). Queda patente cómo, si se considera un rango de valores de k entre 1 y 100, en un principio las prestaciones del sistema mejoran notablemente a medida que aumenta el número de documentos de entrenamiento considerados. Sin embargo, llegado un punto, en este caso para k=22, la clasificación realizada por el sistema cada vez es menos precisa y el

hecho de considerar más documentos de entrenamiento para determinar las categorías empeora los resultados debido a la pérdida de generalidad.

Por otra parte, como se comentó en el diseño del sistema, resulta interesante hacer una reducción de la dimensionalidad del mismo considerando que aquellas palabras que aparecen en muy pocos textos diferentes no serán útiles para llevar a cabo la clasificación. Esto se debe a que palabras muy específicas de cada uno de los textos de entrenamiento, como pueden ser nombres propios, no serán útiles para la posterior comparación con los documentos de set de evaluación y por tanto pueden no ser consideradas a fin de mejorar los resultados.

Para decidir el umbral óptimo, por debajo del cual se considerará que los términos son muy específicos de un documento de entrenamiento dado, se pueden estudiar las prestaciones del sistema a medida que se modifica dicho umbral.

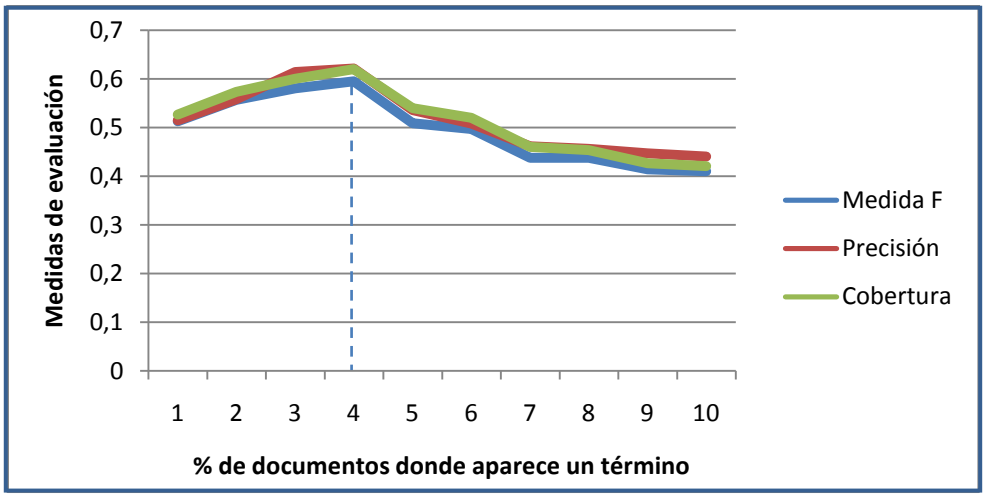

Figura 27. Prestaciones vs umbral para la eliminación de términos

Como se puede observar en la figura anterior (de nuevo estudiando el caso de cinco categorías y para $k=k_{óptima}$), las prestaciones, en un principio, mejoran a medida que aumenta el umbral que indica el número de textos diferentes en los que debe aparecer un término para ser relevante de cara a la clasificación. Una vez superado el umbral del 4% (umbral óptimo) se puede observar que las prestaciones comienzan a empeorar y que, por tanto, de considerar un umbral superior, se estarían eliminando términos que sí son relevantes para clasificar adecuadamente los nuevos documentos del set de evaluación.

Como se explicó en el aparatado que corresponde al diseño del sistema (**apartado 4.2**), se han añadido algunos bloques de cara a mejorar las prestaciones del mismo. Además de determinar la k óptima y de emplear un umbral para eliminar las palabras poco representativas también se ha empleado un archivo de sinónimos y se ha tenido en cuenta la presencia de modificadores de la intensidad de las palabras.

Para demostrar la eficacia de estos bloques implementados se presentan a continuación los resultados de la evaluación del sistema (para $k_{óptima}$=22) haciendo uso o no de dichos bloques. Los resultados que se presentan corresponden al caso más general de tener cinco categorías y considerando una asignación de pesos basada en TF.

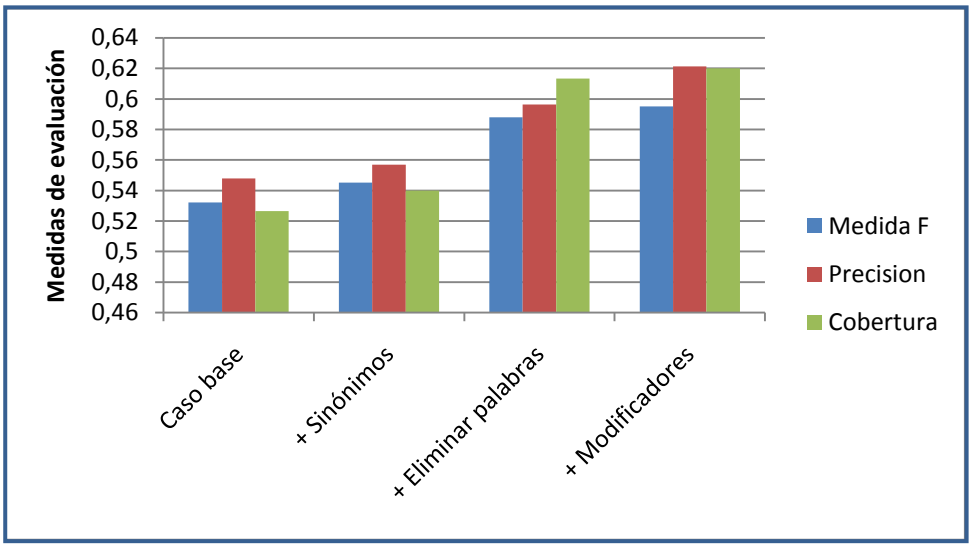

Figura 28. Medidas de evaluación vs bloques añadidos al sistema

Como se puede observar, las prestaciones del sistema mejoran a medida que se añaden a su implementación los distintos bloques diseñados. Al añadir un fichero de sinónimos, que consigue que palabras con el mismo significado en diferentes documentos sean consideradas iguales, los resultados obtenidos mejoran ligeramente. Si además se obtiene un umbral óptimo para eliminar, de los términos considerados durante el entrenamiento, aquellos que aparecen en muy pocos documentos, las prestaciones del clasificador mejoran notablemente. Por último si se añade el hecho de tener en cuenta para el cálculo de la frecuencia de aparición de las palabras, aquellos términos capaces de modificar la intensidad de las mismas los resultados obtenidos tras la evaluación final del sistema, de nuevo, mejoran.

Una vez demostrada la adecuación de los bloques implementados en el sistema y la determinación de los valores óptimos de algunos parámetros (valor de k y del umbral de eliminación) se van a presentar los resultados obtenidos para los distintos modelos de clasificador considerados (dos, cuatro y cinco categorías) de acuerdo al plan de pruebas que se ha presentado en la **Figura 25**.

5.3.1.1 Resultados para cinco categorías

El primero de los casos considerados para su evaluación es el clasificador basado en kNN (con $k_{óptima}$ = 22) capaz de diferenciar entre cinco categorías posibles: excelente, buena, indiferente, mala y pésima.

Como ya se ha comentado, el corpus empleado para este caso consta de 500 críticas de cine, perteneciendo 100 a cada una de las categorías contempladas. Para realizar el entrenamiento del sistema se emplea el 70% de los documentos con lo cual el sistema será evaluado finalmente con un total de 150 críticas, de las cuales 30 pertenecen a cada una de las cinco categorías consideradas.

Las pruebas realizadas pueden dividirse básicamente en dos bloques de acuerdo al algoritmo de asignación de pesos empleado: basado en TF o binario. En la **Figura 29** se representan los resultados finales obtenidos considerando ambos tipos de asignación.

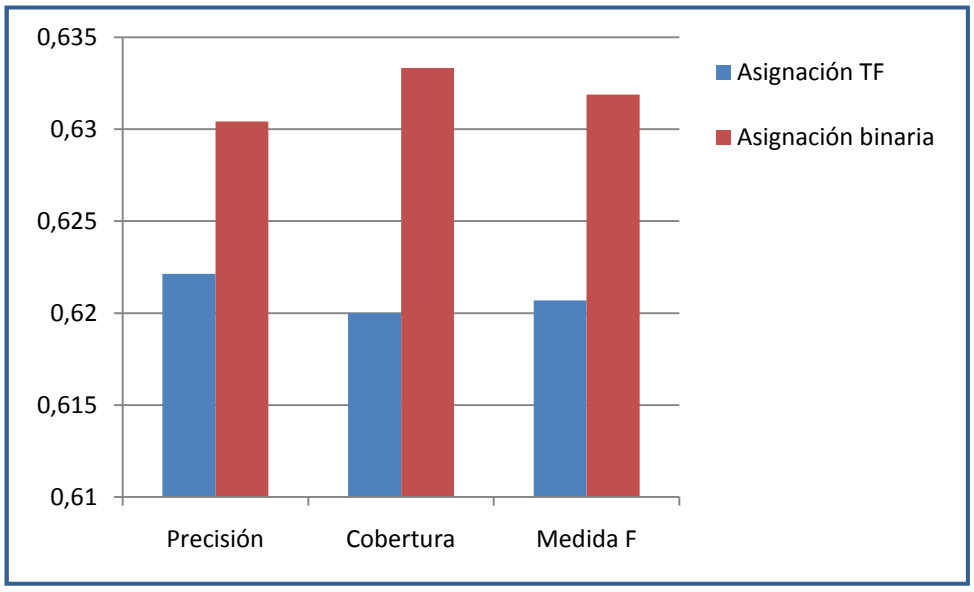

Figura 29. Prestaciones del sistema kNN para cinco categorías

Las prestaciones obtenidas parecen bastante aceptables teniendo en cuenta que puede discernirse entre cinco categorías y que la medida F (que representa conjuntamente la cobertura y la precisión del sistema), en el caso de la asignación binaria, el mejor de los casos, alcanza un valor del 63,2%.

Puede observarse como los resultados obtenidos son más favorables, en el caso de emplear una **asignación de pesos binaria** frente a los obtenidos con una asignación de pesos basada en la frecuencia de aparición de los términos en los documentos (TF). Esto puede deberse a que, a pesar de haber implementado un sistema más complejo (que tiene en cuenta modificadores para la frecuencia de aparición de las palabras), en los documentos que contienen opiniones, como es el caso de las críticas de cine, parece ser más determinante detectar la presencia de algunas palabras clave y no tanto la frecuencia con la que éstas palabras aparecen.

Para analizar con más detalle los resultados obtenidos resulta interesante observar la matriz de confusión del sistema, donde se reflejan todos los crucen entre las categorías del conjunto de test y las asignadas por el sistema. La diagonal de esta matriz se corresponde con los aciertos obtenidos por el sistema para cada una de las categorías. Puesto que los mejores resultados se obtienen para el caso binario son sus resultados los que serán detallados.

Tabla 4. Matriz de confusión sistema kNN con cinco categorías (caso binario)

REAL	Excelentes	Buenas	Indiferentes	Malas	Pésimas
Excelentes	30	0	0	0	0
Buenas	9	16	5	0	0
Indiferentes	1	3	14	11	1
Malas	0	0	3	9	15
Pésimas	0	1	2	2	25

Como se puede observar en la matriz de confusión, las críticas pertenecientes al conjunto de test que son clasificadas más satisfactoriamente son aquellas pertenecientes a las categorías *excelente* (todas las críticas son correctamente clasificadas) y *pésima* (25 de las 30 críticas pertenecientes a ella son correctamente categorizadas). Este hecho no sorprende ya que cabía esperar que los documentos,

tanto del conjunto de entrenamiento como del de test, pertenecientes a las categorías más extremas contuvieran una opinión claramente positiva o negativa lo cual facilitaría la comparación y la posterior identificación de la categoría correcta.

Aunque los documentos pertenecientes a las clases *buena*, *indiferente* y *mala* son clasificados correctamente en menos ocasiones, se puede observar que la confusión se produce con las clases adyacentes, lo cual es un error menos grave ya que la mayoría de los documentos mal clasificados serán asignados a clases no muy alejadas de la verdadera orientación del texto. Por ejemplo, la mayoría de las críticas etiquetadas originalmente como *buenas* y mal clasificadas son asignadas a la categoría *excelente* y la gran mayoría de las críticas cuya verdadera clase es *mala* y son erróneamente categorizadas son asignadas a la categoría *pésima*.

A partir de esta matriz de confusión se puede obtener la tabla de contingencia para cada una de las categorías considerada calculando los parámetros necesarios (TP, FP y FN).

Tabla 5. Tabla de contingencia para cada una de las cinco categorías (caso binario)

	TP	FN	FP
Excelentes	30	0	10
Buenas	16	14	4
Indiferentes	14	16	10
Malas	9	18	13
Pésimas	25	5	16

Como ya se explicó anteriormente, TP_i (verdaderos positivos - *true positives*) representa el total de documentos que han sido correctamente clasificados en la categoría c_i, FN_i (falsos negativos - *false negatives*) es el número de documentos pertenecientes a c_i que no han sido clasificados como tal y FP_i (falsos negativos – *false negatives*) es el número de documentos del conjunto de test que han sido clasificados dentro de c_i cuando su categoría es otra.

Una vez obtenidos estos valores pueden calcularse las medidas de evaluación del sistema (precisión, cobertura y medida F) para cada una de las categorías consideradas de acuerdo a las expresiones presentadas en el **apartado 5.2** para Macro-averaging.

Tabla 6. Medidas de evaluación de las cinco categorías (caso binario)

	PRECISIÓN	COBERTURA	MEDIDA F
Excelentes	0.75	1	0.8571
Buenas	0.8	0.5333	0.64
Indiferentes	0.5833	0.4666	0.5185
Malas	0.4090	0.3333	0.3673
Pésimas	0.6097	0.8333	0.7042

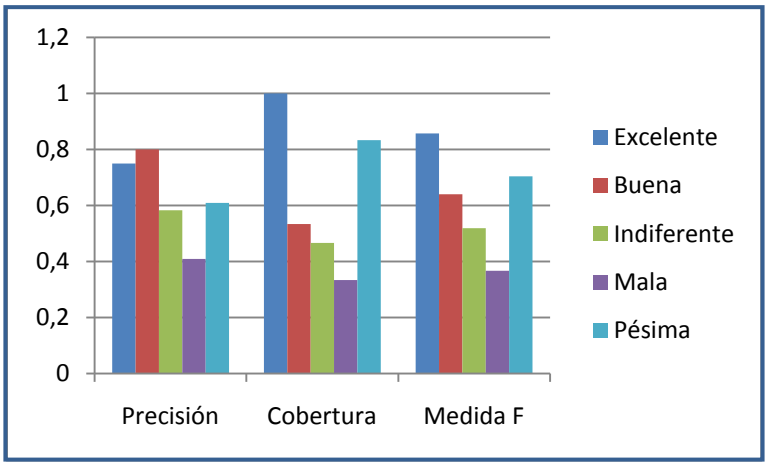

Figura 30. Medidas de evaluación de las cinco categorías (caso binario)

Como cabía esperar tras analizar la matriz de confusión del sistema, las mejores prestaciones se obtienen para la clasificación de los documentos pertenecientes a las clases *excelente* y *pésima* (como puede observarse en la medida F). Por ejemplo, la cobertura es máxima, en el caso de la categoría *excelente* ya que no hay ningún documento de esta clase que sea asignado a otra por error.

Por el contrario, puede afirmarse, a tenor de los resultados, que las categorías que menos fiabilidad presentan son la categoría *mala* e *indiferente*. Esto se debe a que habitualmente una crítica indiferente puede contener ciertos matices de positividad o negatividad que la inclinan ligeramente hacia una de estas dos orientaciones, generalmente hacia la negatividad. Esto empeora los resultados tanto de la cobertura de

las críticas indiferentes (textos originalmente de esta categoría son asignados a otra), como de la precisión de las críticas malas (documentos pertenecientes a otras categorías son asignados a la clase *mala* erróneamente). Por otro lado se puede observar como la cobertura de la categoría *mala* es muy baja y esto se debe a que, como ya se observó en la matriz de confusión, muchas de las críticas que verdaderamente pertenecen a esta clase son asignadas a otra, concretamente a la categoría *pésima* (empeorando por otra parte la precisión de esta última clase).

5.3.1.2 Resultados para cuatro categorías

A fin de poder comparar más adelante con las prestaciones del sistema basado en el uso del diccionario afectivo (que no puede diferenciar documentos de la clase *indiferente*), y visto que los resultados obtenidos en el apartado anterior para la categoría indiferente son de los más negativos, se va a evaluar el sistema basado en kNN teniendo en cuenta únicamente cuatro categorías posibles (*excelente*, *buena*, *mala* y *pésima*).

Para llevar a cabo esta evaluación, como ya se ha comentado, el corpus empleado será básicamente el mismo que en el caso de cinco categorías con la salvedad de que las críticas indiferentes no serán tenidas en cuenta. El corpus, por tanto, constará de 400 críticas, de las cuales 100 pertenecen a cada una de las cuatro categorías contempladas. De nuevo, el 70% de ellas son empleadas para el entrenamiento del sistema con lo que se dispone de 30 críticas de cada clase para realizar la evaluación.

Al igual que se analizó en el caso anterior, se presentan (**Figura 31**) los resultados globales obtenidos por el clasificador en el caso de considerar la asignación de pesos binaria y la asignación TF.

Se puede observar cómo, de nuevo, queda patente que la asignación de pesos binaria, que detecta la presencia o no de ciertos términos, es más apropiada para la clasificación de este tipo de documentos. En el mejor de los casos (el caso binario) la medida F del sistema es superior al 70% con lo que podemos decir que los resultados para cuatro categorías son aceptables.

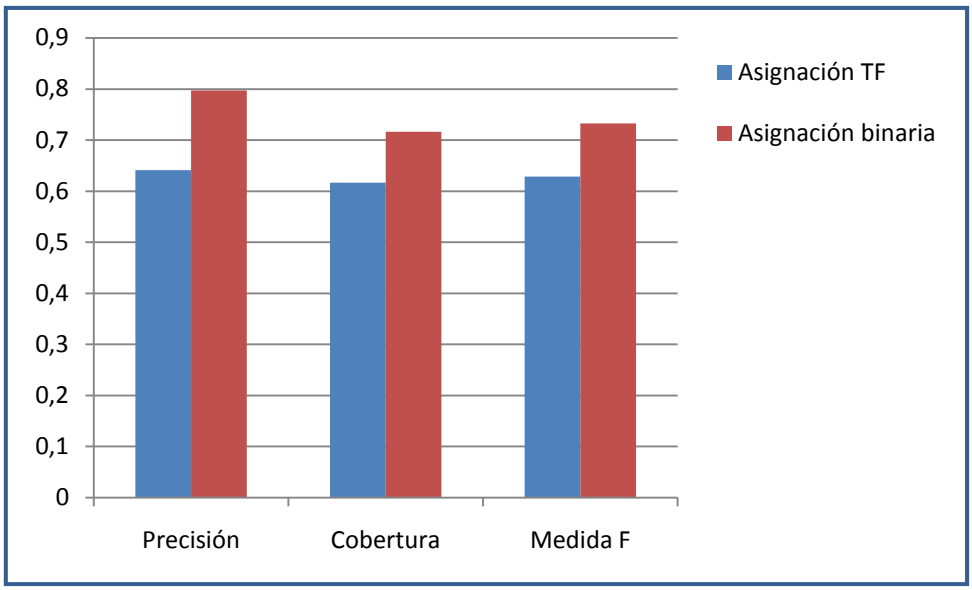

Figura 31. Prestaciones del sistema kNN con cuatro categorías

Para analizar mejor los resultados de la clasificación, a continuación se presenta la matriz de confusión del sistema para el caso binario donde aparecen todos los cruces entre las categorías reales del set de evaluación y las obtenidas por el clasificador.

Tabla 7. Matriz de confusión sistema kNN con cuatro categorías (caso binario)

REAL	Excelentes	Buenas	Malas	Pésimas
Excelentes	29	1	0	0
Buenas	9	17	2	2
Malas	2	0	14	14
Pésimas	1	1	2	26

Como ocurría en el caso anterior, para cinco categorías, los documentos de test que son asignados por el clasificador con mayor acierto son aquellos pertenecientes a las categorías extremas, es decir, a las clases *excelente* y *pésima* (acertando en la práctica totalidad de los casos). En el caso de las categorías *buena y mala* se acierta únicamente en, aproximadamente, la mitad de los casos aunque si cabe destacar que los fallos se producen fundamentalmente con las categorías más cercanas, *excelente* y *pésima* respectivamente, lo que hace que el error sea menos relevante ya que al menos se detecta la orientación global del texto.

A partir de esta matriz de confusión se puede obtener la tabla de contingencia para cada una de las cuatro categorías consideradas calculando los parámetros necesarios (TP, FP y FN).

Tabla 8. Tabla de contingencia para cada una de las cuatro categorías (caso binario)

	TP	FN	FP
Excelentes	29	1	12
Buenas	17	13	2
Malas	14	16	4
Pésimas	26	4	16

Una vez obtenidos estos valores pueden calcularse las medidas de evaluación del sistema para cada una de las categorías consideradas.

Tabla 9. Medidas de evaluación de las cuatro categorías (caso binario)

	PRECISIÓN	COBERTURA	MEDIDA F
Excelentes	0.7073	0.9666	0.8169
Buenas	0.8947	0.5666	0.6938
Malas	0.7777	0.4666	0.5833
Pésimas	0.6190	0.8666	0.7222

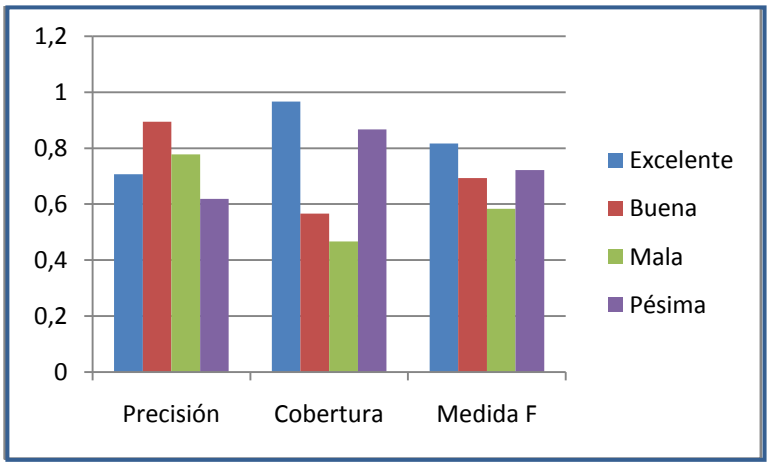

Figura 32. Medidas de evaluación de las cuatro categorías (caso binario)

A tenor de los resultados obtenidos se demuestra, como ya se sabía, que las categorías *excelente* y *pésima* presentan unas prestaciones muy superiores con respecto a las clases *buena* y *mala* (véanse los resultados de la medida F) y por tanto documentos de set de evaluación pertenecientes a estas dos categorías serán clasificados correctamente en un mayor número de ocasiones. La categoría *mala* es con la que peores resultados se obtienen y sus prestaciones y fundamentalmente su cobertura es muy baja. Esto mismo ocurre en el caso de la categoría *buena y* se debe a que, como se pudo observar en la matriz de confusión, muchas de las críticas pertenecientes a estas clases son categorizadas como pésimas y excelentes respectivamente (lo cual por otra parte no es un error demasiado grave) lo que además conlleva un notable empeoramiento en la precisión de esta dos últimas categorías (la precisión asociada a la clases *pésima* y *excelente* es la más baja).

5.3.1.3 Resultados para dos categorías

Como última prueba para evaluar el sistema basado en kNN se ha considerado el caso de obtener un clasificador capaz de diferenciar entre dos únicas categorías, *positiva* y *negativa*. Se trata, por tanto, de evaluar un clasificador binario de opinión capaz de decidir la orientación afectiva de las críticas del set de evaluación.

Para llevar a cabo la evaluación, el corpus empleado es exactamente el mismo que en el caso anterior (considerando cuatro categorías) con la particularidad de que las críticas excelentes y buenas serán englobadas dentro de la clase *positiva* y las críticas malas y pésimas serán englobadas dentro de la clase *negativa*. De esta forma, el corpus empleado consta de un total de 400 críticas, de las cuales 200 pertenecen a cada una de las dos categorías. Como en los casos anteriores, el 70% de los documentos serán empleados para el entrenamiento de sistema y por tanto la evaluación del mismo se llevará a cabo con 120 críticas de las cuales 60 son positivas y 60 negativas.

Siguiendo los pasos llevados a cabo en los dos apartados anteriores, se presentan a continuación (**Figura 33**) los resultados globales obtenidos por el sistema tanto si la asignación de pesos es binaria como si está basada en la frecuencia de aparición de los términos.

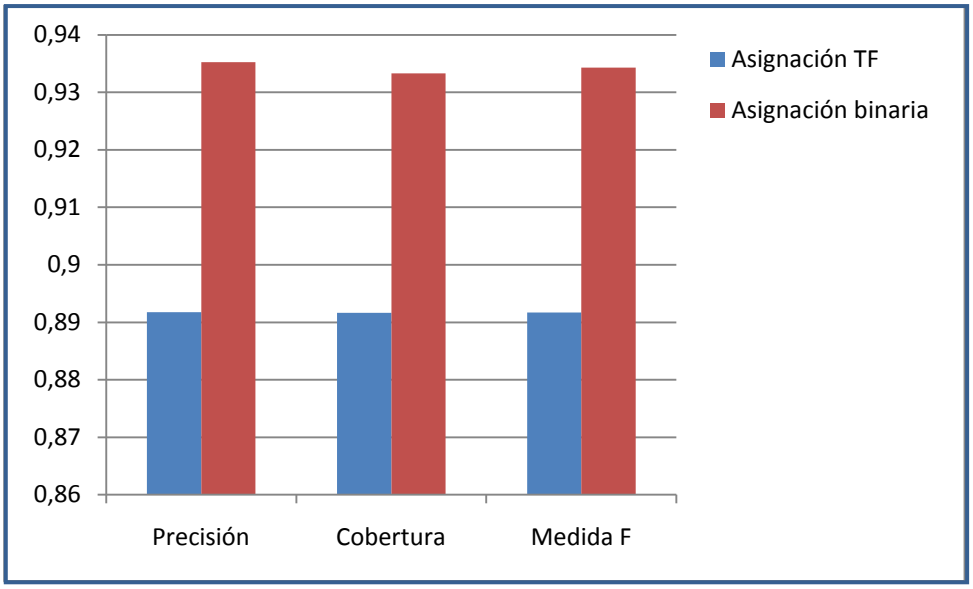

Figura 33. Prestaciones del sistema kNN con dos categorías

Se puede observar cómo, al igual que ocurría en los casos anteriormente evaluados, la asignación de pesos binaria, basada en la presencia o no de los términos en los documentos, proporciona unos resultados más satisfactorios a la hora de obtener la categoría de las críticas pertenecientes al set de evaluación frente a la asignación TF. Las prestaciones del sistema en el caso de diferenciar entre críticas cuya orientación es positiva o negativa son muy elevadas y se alcanza una medida de F para el caso binario de más del 93%.

A continuación se presenta la matriz de confusión para el caso de asignación de pesos binaria, que es el diseño que mejores prestaciones presenta.

Tabla 10. Matriz de confusión sistema kNN con dos categorías (caso binario)

REAL	Positiva	Negativa
Positivas	58	2
Negativa	6	54

Como cabía esperar una vez conocidos los resultados de la evaluación del sistema, se puede observar que prácticamente la totalidad de las críticas del conjunto de test son clasificadas correctamente. Este hecho podía suponerse si se tienen en cuenta los resultados obtenidos para el caso de cuatro categorías ya que la gran mayoría de los

errores en la clasificación en ese caso se producían con las categorías más cercanas. Esto quiere decir que, por ejemplo, si una crítica originalmente está etiquetada como excelente o buena y el sistema determina que la clase resultado es *positiva* no se tiene en cuenta si se ha acertado exactamente con el grado de positividad, simplemente si la orientación determinada es correcta. Al aumentar la generalidad del sistema y reducir la exigencia en cuanto a la determinación de la intensidad de la orientación afectiva del texto parece lógico que los resultados finales sean más satisfactorios.

A continuación se muestra la tabla de contingencia, obtenida a partir de la matriz de confusión del sistema, donde se presenta el valor de los parámetros TP, FP y FN.

Tabla 11. Tabla de contingencia para cada una de las dos categorías (caso binario)

	TP	FN	FP
Positiva	58	2	6
Negativa	54	6	2

A partir de los valores presentados en la tabla anterior, se pueden calcular las medidas de evaluación del sistema (precisión, cobertura y medida F) particularizadas para cada una de las dos categorías consideradas.

Tabla 12. Medidas de evaluación de las dos categorías (caso binario)

	PRECISIÓN	COBERTURA	MEDIDA F
Positiva	0.9062	0.9666	0.9354
Negativa	0.9642	0.9	0.9310

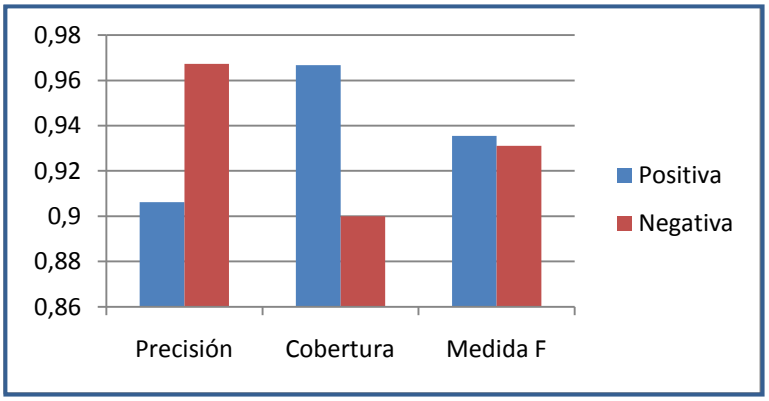

Figura 34. Medidas de evaluación de las dos categorías (caso binario)

Las prestaciones obtenidas para ambas categorías son muy similares y muy satisfactorias (la medida F es superior al 93% en ambos casos). Como puede observarse la precisión es inferior en el caso de las críticas positivas debido a que más documentos negativos son clasificados erróneamente como positivos que lo que ocurre en el caso contrario. De forma complementaria, la cobertura es superior en el caso de las críticas positivas ya que menos documentos pertenecientes a esta categoría son clasificados por error como pertenecientes a la clase contraria.

El hecho de que más críticas negativas sean clasificadas como positivas por el sistema se debe, posiblemente, al efecto de la negación y la ironía, aspectos que no se están tratando. A continuación se muestran varios ejemplos de críticas negativas que son clasificadas erróneamente como positivas:

Fui a cine con grades expectativas porque las películas de este director siempre me han encantado y cuando apenas habian transcurrido los diez primeros minutos ya sabía que aquella no me iba a gustar. El reparto es muy bueno, digno de una gran película, pero el argumento no consigue enganchar en ningun momento. La verdad es que no la recomiendo.

Esta película es el último título de la factoría Almodovar, el que se supone es el mejor director del cine español. Otra vez nos sorprende Almodovar con un argumento de esos que la gente alaba porque es suyo, tan tan original y con unos personajes tan de la vida cotidiana que siempre que termino de ver sus películas salgo con la sensación de que me gustan tanto como una patada en la espinilla. Qué gran director…estoy deseando que se ponga a grabar la siguiente, sí sí.

5.3.2. SISTEMA BASADO EN EL EMPLEO DE UN DICCIONARIO AFECTIVO

En segundo lugar, se presentan y analizan los resultados obtenidos para el sistema implementado basado en el empleo de un diccionario afectivo. En esta ocasión, las pruebas se dividirán únicamente en dos grandes bloques, que corresponden a la clasificación de acuerdo a dos categorías (*positiva* y *negativa*). En este caso el análisis del sistema considerando cinco categorías no ha sido posible ya que la determinación de un umbral adecuado para las críticas indiferentes no era viable con el contenido del diccionario empleado, que únicamente asigna una orientación positiva o negativa a los términos que contiene. Las críticas indiferentes no son únicamente aquellas situadas entre las positivas y negativas si no que disponen de un vocabulario propio difícil de etiquetar con el uso del diccionario de que se dispone.

5.3.2.1 Resultados para dos categorías

En primer lugar, con el diseño inicial de este sistema se buscaba que fuera capaz de distinguir entre críticas cuya orientación fuera positiva o negativa. Para conseguirlo (como se explicó en el **apartado ¡Error! No se encuentra el origen de la referencia.**), ebía establecerse el umbral a partir del cual valoración afectiva de los documentos obtenida del diccionario (debida a la orientación de cada término) haría que estos fueran considerados positivos o negativos.

El corpus empleado para la evaluación de este sistema es el mismo que el empleado, y anteriormente explicado, en el caso del sistema de clasificación binario (dos categorías) basado en el algoritmo kNN.

Puesto que únicamente debían sumarse términos positivos y negativos de cada documento y establecer un valor resultante, el primer umbral considerado para la clasificación fue cero **(U_{PosNeg} = 0).** A continuación se muestran los resultados obtenidos considerando este umbral de decisión y sin la necesidad de realizar una fase de entrenamiento previa del sistema.

Tabla 13. Matriz de confusión clasificación binaria con U=0

REAL	Positivas	Negativas
Positivas	60	0
Negativas	20	40

Como se puede observar en la matriz de confusión del sistema, todas las críticas positivas del conjunto de test son clasificadas correctamente mientras que las negativas son en muchos casos (la tercera parte de las veces) clasificadas como positivas por el sistema. Esto hace sospechar que el umbral de decisión no es apropiado y debe desplazarse hacia el lado positivo con el fin de conseguir que más críticas negativas sean clasificadas adecuadamente.

A continuación se presentan las prestaciones finales proporcionadas por el sistema tras obtener la matriz de contingencia y calcular la precisión, cobertura y medida F para las dos categorías.

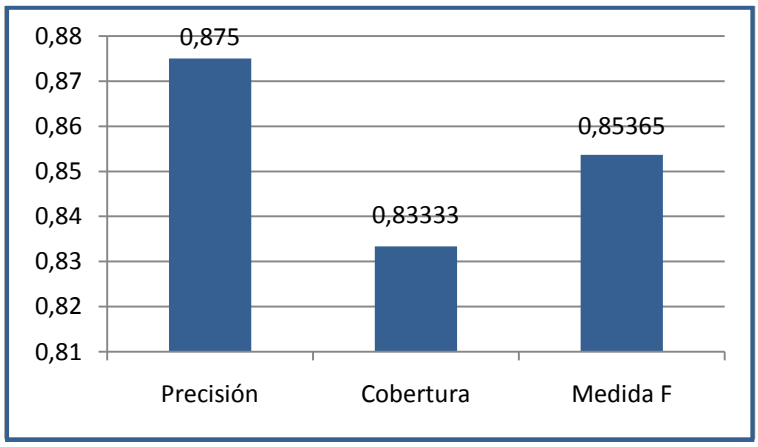

Figura 35. Medidas de evaluación del sistema con dos categorías (U=0)

Como ya se ha comentado, tras observar la matriz de confusión del sistema se intuye que el umbral de decisión óptimo para determinar la orientación afectiva de los documentos no es cero y debe ser un valor positivo. Por ello se llevó a cabo el entrenamiento del sistema a fin de determinar el umbral para realizar la clasificación.

Dicho umbral es el valor medio entre la media de la valoración afectiva global de los documentos positivos y negativos del set de entrenamiento (ver **apartado 4.3.2.2**).

El umbral óptimo (normalizado) determinado tras la fase de entrenamiento es:

$$U_{PosNeg} = U_{medio} = 0.0241$$

A continuación se presentan los resultados obtenidos tras emplear para la clasificación el umbral anteriormente determinado.

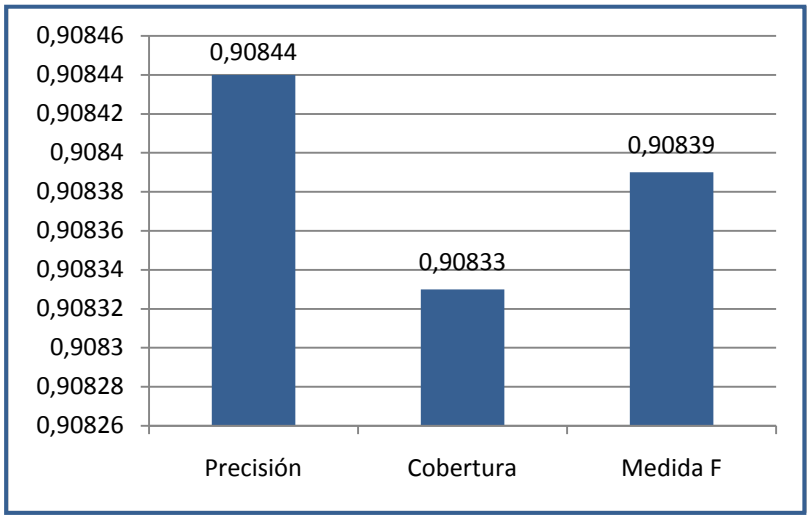

Figura 36. Medidas de evaluación del sistema ($U = U_{medio}$)

Se puede observar que los resultados globales han mejorado con respecto al caso anterior, donde el umbral estaba situado en el valor cero, y ahora las prestaciones del sistema son más satisfactorias ya que se alcanza una medida F por encima del 90%.

Para analizar con más detalle los resultados obtenidos tras la utilización de este umbral, a continuación se muestra la matriz de confusión del sistema.

Tabla 14. Matriz de confusión clasificación binaria ($U_{PosNeg}=U_{medio}$)

REAL	Positivas	Negativas
Positivas	55	5
Negativas	6	54

Se puede observar que, en esta ocasión, la gran mayoría de las críticas positivas y negativas del conjunto de test son clasificadas correctamente y ninguna de las dos clases presenta unos resultados claramente mejores o peores con respecto a la otra. Esto no ocurría en el caso anterior ($U_{PosNeg} = 0$) donde las críticas positivas eran siempre correctamente clasificadas y sin embargo las críticas negativas no lo eran en multitud de ocasiones.

A continuación se muestra la tabla de contingencia, obtenida a partir de la matriz de confusión del sistema, donde se presenta el valor de los parámetros TP, FP y FN.

Tabla 15. Tabla de contingencia para cada una de las dos categorías

	TP	FN	FP
Positivas	55	5	6
Negativas	54	6	5

A partir de los valores presentados en la tabla anterior, se pueden calcular las medidas de evaluación del sistema particularizadas para cada una de las dos categorías consideradas.

Tabla 16. Medidas de evaluación para cada una de las dos categorías

	PRECISIÓN	COBERTURA	MEDIDA F
Positivas	0.9016	0.9166	0.9090
Negativas	0.9152	0.9	0.9075

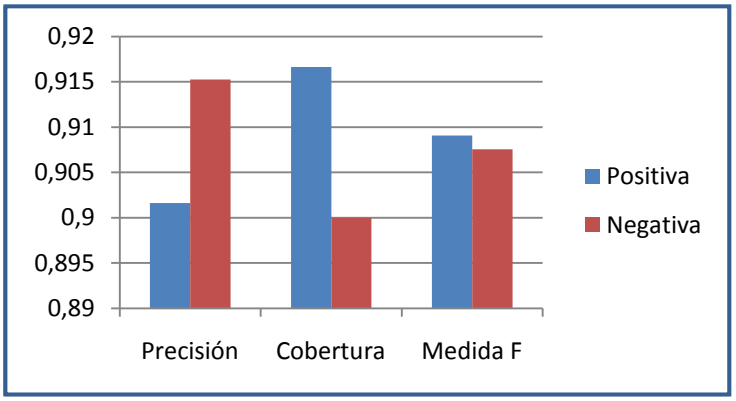

Figura 37. Medidas de evaluación para las dos categorías

Como se puede observar las prestaciones del sistema a la hora de clasificar adecuadamente críticas positivas y negativas son, además de muy satisfactorias, muy similares. Esto indica que el umbral determinado gracias al entrenamiento del sistema parece adecuado ya que se consigue un equilibrio en la capacidad del sistema para categorizar correctamente documentos de una u otra orientación afectiva (cosa que no ocurría en el caso anterior donde el sistema era capaz fundamentalmente de clasificar adecuadamente críticas positivas).

5.3.2.2 Resultados para cuatro categorías

La segunda de las pruebas realizadas para el clasificador basado en el empleo del diccionario afectivo consiste en evaluar un clasificador que, al igual que el implementado basándose en el algoritmo kNN, sea capaz de distinguir entre cuatro categorías posibles (*excelente*, *buena*, *mala* y *pésima*).

Para llevar a cabo la evaluación de este sistema el corpus empleado será el ya comentado para el sistema de cuatro categorías basado en kNN.

Visto que para el sistema anterior (dos categorías) la determinación del umbral entrenando el sistema produce resultados satisfactorios, la primera prueba realizada consiste en obtener las prestaciones del sistema si los umbrales de decisión (en este caso tres) son obtenidos teniendo en cuenta la **orientación afectiva** media de cada una de las clases consideradas dentro del conjunto de entrenamiento.

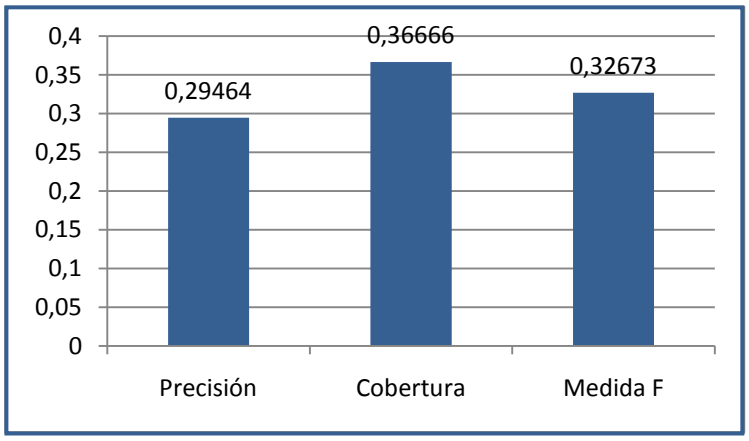

Figura 38. Medidas de evaluación para las cuatro categorías

Los resultados muestran que las prestaciones del sistema no son muy favorables ya que, por ejemplo, apenas se alcanza para la medida F el 30%. Para analizar en detalle la causa de estos resultados se observa la matriz de confusión de sistema:

Tabla 17. Matriz de confusión del sistema con cuatro categorías

REAL	Excelentes	Buenas	Malas	Pésimas
Excelentes	8	22	0	0
Buenas	18	12	0	0
Malas	0	17	0	13
Pésimas	0	3	3	24

De acuerdo al contenido de la matriz de confusión, se puede afirmar que las únicas críticas que parece clasificar adecuadamente el sistema son las pésimas. El resto de documentos pertenecientes a las otras categorías serán clasificados erróneamente en la gran mayoría de los casos. Un caso significativo parece producirse entre las categorías *excelente* y *buena*, ya que la mayor parte de las críticas excelentes son clasificadas por el sistema como buenas y lo mismo ocurre a la inversa, la mayor parte de las críticas buenas son clasificadas como excelentes.

A tenor de los resultados, los umbrales obtenidos teniendo en cuenta únicamente el valor de la orientación afectiva media de cada una de las categorías consideradas durante el entrenamiento, no parecen ser adecuados para implementar un sistema efectivo capaz de distinguir entre las cuatro categorías definidas.

Como alternativa se decide evaluar un sistema cuyos umbrales de decisión son determinados no sólo teniendo en cuenta el valor de la orientación de los términos dentro del diccionario (si son positivos o negativos) si no también la **intensidad** de los mismos.

Se ha demostrado que el umbral obtenido en el caso del sistema anterior (dos categorías) es efectivo a la hora de diferenciar la orientación afectiva de los documentos, es decir determinar si son positivos o negativos, así que se empleará este umbral para distinguir, en primer lugar, la orientación de los documentos del conjunto

de test. Una vez determinada, para poder diferenciar el grado de positividad (excelente o buena) o negatividad (mala o pésima) de las críticas se obtendrán los umbrales a partir del valor medio de intensidad afectiva (en vez de orientación) obtenido para las distintas categorías durante el entrenamiento del sistema.

A continuación se muestran los resultados obtenidos empleando estos nuevos umbrales para la clasificación.

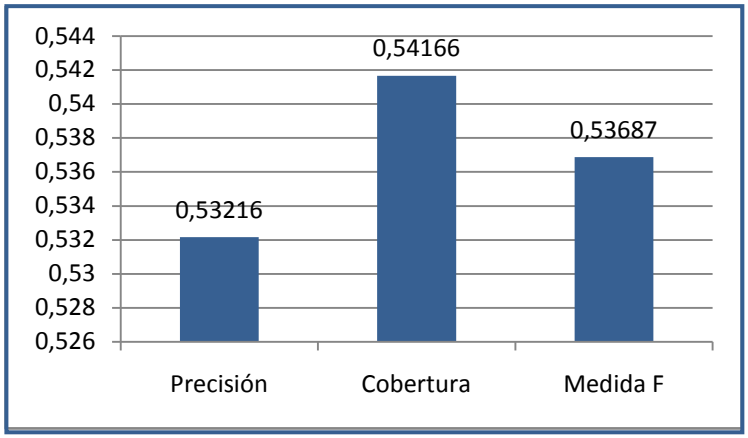

Figura 39. Medidas de evaluación con cuatro categorías

Se puede observar que los resultados han mejorado sensiblemente con respecto al caso anterior en que para determinar los umbrales sólo se tenía en cuenta el valor medio de la orientación de las categorías del conjunto de entrenamiento. Al discernir el grado de positividad y negatividad a través del valor de la intensidad afectiva de cada categoría se obtienen unas prestaciones mucho más favorables para el sistema, alcanzado la medida F un valor cercano al 54%.

Para analizar con detalle los resultados obtenidos tras la evaluación del sistema se muestra a continuación la matriz de confusión del mismo.

Tabla 18. Matriz de confusión sistema con cuatro categorías

REAL	Excelentes	Buenas	Malas	Pésimas
Excelentes	18	9	3	0
Buenas	16	12	2	0
Malas	2	4	11	13
Pésimas	0	0	6	24

Se puede observar que el número de documentos correctamente clasificados para todas las categorías aumenta considerablemente con respecto al caso anterior. Además, otro aspecto positivo es que los errores que se producen en la clasificación son en su gran mayoría con las clases adyacentes, es decir con aquella a las que más se parecen, con lo que dichos errores podrían considerarse menos importantes.

Para finalizar se presenta la tabla de contingencia para todas las categorías a parir de la cual se pueden obtener las medidas de evaluación (precisión, cobertura y medida F) concretas para cada una de las clases.

Tabla 19. Tabla de contingencia para cada una de las cuatro categorías

	TP	FN	FP
Excelentes	18	12	18
Buenas	12	18	13
Malas	11	19	11
Pésimas	24	6	13

Tabla 20. Medidas de evaluación para cada una de las cuatro categorías

	PRECISIÓN	COBERTURA	MEDIDA F
Excelentes	0.5	0.6	0.5454
Buenas	0.48	0.4	0.4363
Malas	0.5	0.3666	0.4230
Pésimas	0.6486	0.8	0.7164

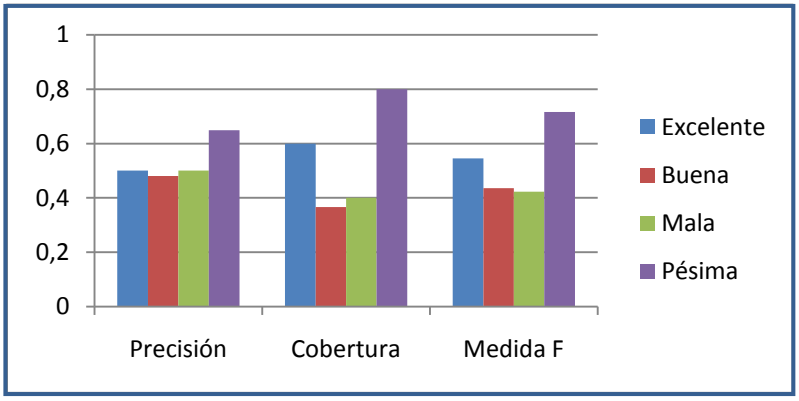

Figura 40. Medidas de evaluación para las cuatro categorías

Se puede observar como las prestaciones del sistema son bastante aceptables y varían notablemente en función de la categoría del documento del set de evaluación a clasificar. Como se ha observado al evaluar también otros sistemas, las categorías que mejores resultados ofrecen, como cabía esperar, son las más extremas, en este caso *excelente* y *pésima,* ya que, en general, se trata de críticas que expresan de una forma muy rotunda e inequívoca la orientación afectiva de las mismas.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. CONCLUSIONES

En este trabajo se presenta un primer paso para la implementación de un sistema de detección automática de emociones de críticas de cine en español, capaz de determinar la orientación afectiva de las mismas.

Se ha optado por emplear dos métodos para su implementación: uno basado en el algoritmo de los k vecinos más próximos (kNN-K Nearest Neighbour) y otro basado en el empleo de un diccionario afectivo, para así poder obtener las prestaciones del sistema haciendo uso de dos mecanismos muy diferentes.

Para el primero de los sistemas desarrollados puede decirse que, efectivamente, el empleo del **algoritmo kNN** aporta sencillez y eficacia a la implementación y evaluación del mismo. Los resultados obtenidos tras la evaluación de este primer sistema, muestran la importancia de determinar el valor óptimo de la k (número de documentos del set de entrenamiento con los que se comparará cada documento del conjunto de test) ya que las prestaciones del sistema varían notablemente según el número de *vecinos* considerados.

Al añadir los distintos bloques diseñados para mejorar el sistema se comprueba que las prestaciones del mismo mejoran notablemente. Esto permite afirma que el empleo de un diccionario de sinónimos (para que palabras con el mismo significado en diferentes textos sean consideradas iguales), la eliminación de los términos que aparecen en muy pocos textos diferentes y la consideración de los modificadores que intensifican o disminuyen la importancia de los términos dentro de un texto, son mecanismos que aportan valor al sistema y mejoran sensiblemente sus prestaciones.

La evaluación de este sistema se ha llevado a cabo realizando pruebas para considerando cinco, cuatro y dos categorías y, en todos los casos se puede afirmar que el método de asignación de pesos que mejores resultados ofrece es el binario frente al que se basa en la frecuencia de aparición de los términos. Esto es, vectores binarios de características donde las entradas indican si un término aparece (valor =1) o no (valor =0) formaron una base más efectiva para la clasificación de la polaridad de críticas, frente a vectores de características donde las entradas se valoraban según la frecuencia de aparición de los términos. Este hallazgo puede indicar una importante diferencia entre la clasificación afectiva y la categorización típica de textos basada en temas. Mientras que un tema es más probable que sea enfatizado por la aparición frecuente de ciertas palabras clave, el sentimiento en general no suele ser realzado a través del uso repetido de los mismos términos.

A modo de resumen se puede decir que los resultados obtenidos tras la evaluación del sistema basado en kNN son bastante satisfactorios obteniendo una medida F (qué combina la precisión y la cobertura del sistema) del 63.2% para el clasificador de cinco categorías, del 73,48% para cuatro categorías posibles y del 93,42% en el caso de diferenciar sólo entre críticas positivas y negativas.

El segundo de los sistemas evaluados es el basado en el empleo de un **diccionario afectivo**. En este caso se comprobó que únicamente haciendo uso de las valoraciones de las palabras contenidas en el diccionario, no es posible implementar el sistema capaz de identificar críticas indiferentes con lo que se obtuvieron las prestaciones del sistema considerando únicamente el caso de distinguir entre cuatro categorías (*excelente*, *buena*, *mala* y *pésima*) y dos categorías (*positiva* y *negativa*).

Los umbrales óptimos de decisión se obtienen gracias al entrenamiento del sistema. Con el umbral determinado para el caso de dos categorías, de acuerdo a la valoración global de la orientación afectiva de cada una de las dos clases durante el entrenamiento, las prestaciones obtenidas son bastante satisfactorias, alcanzando la precisión, la cobertura y la medida F del sistema un valor por encima del 90%. En el caso de evaluar el sistema que debe ser capaz de discernir entre cuatro categorías los valores más favorables se obtienen si, en primer lugar, se determina la orientación (positiva o negativa) de las críticas de la misma forma que se hace en el caso de tener sólo dos clases y a continuación se emplea la intensidad afectiva media de cada una de las clases

durante el entrenamiento para determinar exactamente el grado de positividad o negatividad de los documentos. Con los umbrales de decisión determinados de esta manera, las prestaciones del sistema son tales que la precisión es del 53.2%, la cobertura del 54.1% y la medida F del sistema alcanza el 53.6%.

De estos resultados se puede deducir que, si bien para el caso de la clasificación binaria (críticas positivas y negativas) las prestaciones del sistema basado en kNN y el basado en el diccionario afectivo son muy similares (por encima del 90%), el sistema que emplea kNN resulta bastante más efectivo (medida F del 73,48% frente al 53.6% conseguido con el diccionario) a la hora de concretar el grado de positividad y negatividad de las críticas y discernir así entre opiniones excelentes o buenas y malas o pésimas, algo para lo que el empleo del diccionario utilizado no parece ser suficiente.

Se ha podido observar, por otra parte, tras la evaluación de los dos sistemas que, en el caso de tener más de dos categorías, las críticas que siempre son mejor clasificadas, con una tasa de acierto muy elevada, son aquellas pertenecientes a las categorías extremas, es decir las críticas excelentes y pésimas. Esto se debe a que el lenguaje empleado y las opiniones vertidas son mucho más rotundas y hacen que sea más difícil errar en la determinación de la clase a la que pertenecen.

Tras el análisis de todos los resultados obtenidos se puede concluir que, con el corpus empleado, los sistemas implementados clasifican más o menos satisfactoriamente las críticas del conjunto de evaluación. No obstante no se debe perder de vista que aunque el corpus empleado ha sido revisado y adecuado para este estudio, en general, las críticas vertidas por usuarios pueden suponer un gran problema a la hora de llevar a cabo la clasificación de opinión ya que los textos suelen ser informales con lo que la calidad de los mismos, las faltas de ortografía y el vocabulario empleado puede dificultar enormemente este tipo de tareas de clasificación.

La realización de este proyecto constituye sólo una pincelada en el análisis de la clasificación automática de documentos de opinión en español, más concretamente críticas de cine, y deja abierto un amplio camino para una investigación más profunda relacionada con el campo de la clasificación afectiva de textos. Este tema se ha convertido en una actual y activa área de investigación debido en gran parte a la variedad de posibles aplicaciones que proporciona, que van desde el análisis

automatizado de opiniones de películas, obras, o productos en general, hasta estudios que permitan dar seguimiento a cómo va evolucionando la percepción de los ciudadanos acerca de las empresas, gobernantes o políticos.

6.2. TRABAJOS FUTUROS

Existen numerosas líneas de trabajo que surgen como ideas a lo largo de la implementación de este proyecto, y sería interesante seguir investigando sobre ellas a fin de mejorar las prestaciones del sistema. A continuación se detallan algunas de estas posibles futuras líneas de investigación.

- Emplear nuevos algoritmos de aprendizaje (aparte del kNN) así como diferentes métodos para la asignación de pesos (por ejemplo TF*IDF) a fin de poder comparar los resultados y aplicar aquél que proporcione mejores prestaciones.

- Emplear otros diccionarios afectivos existentes (por ejemplo variantes del *SentimentWordNet*) para determinar la influencia de la elección del mismo en los resultados obtenidos por este tipo de sistemas.

- Abordar el tratamiento de la negación en análisis de las críticas para detectar correctamente la orientación afectiva de las mismas. No existe una situación paralela en la extracción de información clásica, donde un sólo término de negación juegue un papel tan importante en la clasificación de documentos de opinión ("*no me gusta esta película*"), por tanto sería interesante tratar de abordar esta tarea. No obstante debe tenerse en cuenta que existe una gran dificultad en el modelado de la negación ya que ésta se puede expresar de formas muy sutiles y, por ejemplo, el sarcasmo y la ironía pueden ser bastante difíciles de detectar ("*la película me gusta mucho, tanto como una patada en la espinilla*").

- Tratar de determinar la subjetividad de un documento para así decidir si la naturaleza de un texto dado atiende a hechos (describen una situación dada o un hecho, sin expresar una opinión positiva o negativa en él) o expresa una opinión. En este proyecto se ha asumido que las críticas empleadas son

puramente de opinión, sin embargo en muchas ocasiones la crítica suele ir acompañada de una descripción objetiva del argumento de la película o de la filmografía de los actores con lo que identificar y trabajar sólo con aquellas porciones de los documentos que son subjetivas, y por tanto expresan opinión, tendrá presumiblemente un impacto positivo en la determinación de la orientación de la crítica.

- Otra posible línea de investigación relacionada puede ser conseguir analizar el lenguaje extraído de fragmentos de audio (como pueden ser las locuciones en contestadores automáticos de un servicio de atención al cliente o de encuestas telefónicas) a fin de determinar la orientación afectiva de las opiniones vertidas, empleando básicamente los mismos mecanismos aplicados en este proyecto.

ANEXO A – FICHERO DE SINÓNIMOS

excelente perfecto emocionante maravilloso genial impresionante fantástico espectacular formidable espléndido excepcional magnífico estupendo fabuloso joya extraordinario superior prodigioso soberbio fascinante tremendo magistral apasionante inolvidable precioso sobresaliente admirable memorable impecable;

encantar maravillar fascinar deleitar embelesar embrujar impresionar apasionar;

agradar amenizar distraer entretener animar divertir;

original sorprendente ocurrente nuevo insólito inédito singular curioso asombroso;

gustar satisfacer disfrutar contentar alegrar;

contento satisfecho alegre;

animado alocado divertido;

agradable ameno grato interesante distraído satisfactorio entretenido entrañable;

pésimo patético mierda horrible lamentable nefasto deprimente pavoroso terrible espantoso horripilante horroroso fatal detestable lastimoso repelente insoportable soporífero desesperante insufrible inaguantable;

absurdo tonto raro;

malo flojo cutre decepcionante pobre vacío mediocre;

absurdo tonto raro;

aburrir cansar empalagar molestar desanimar fastidiar agobiar cargar hastiar hartar;

decepcionar desencantar defraudar;

aburrido inapetente cansado latoso lento desanimado incómodo molesto pesado soporífero cargante tedioso cansino interminable tostón;

decepcionar defraudar desilusionar desengañar;

disgustar desagradar;

ANEXO B – FICHEROS DE MODIFICADORES

2 súper totalmente absolutamente;

1 muy;

0.5 bastante;

-0.5 poco un_poco demasiado;

-0.9 nada;

Modificadores de adjetivos

2 muchísimo;

1 mucho;

0.5 bastante;

-0.5 poco demasiado;

-0.8 poquísimo;

-0.9 nada;

-0.9 nada_de_nada;

Modificadores de verbos

ANEXO C –DICCIONARIO AFECTIVO

flojo N;pretencioso N;homofóbico N+;abusivo N;escatológico N;agotador N+;interesante P;lejos -;original P;llamativo +;ramplonería N+;marchito N;desangelado N;vergonzosa N;zafia +;soso N;artesano P;profesional P+;desamparar N-;abandonar N-;abyecto N;capaz P+;anodino N;anormal N;abominable N+;abundar P;abrasivo N+;raspante N+;brusco N;ausencia N-;falta N-;falto N-;faltar N-;absoluta +;absoluto +;absurdo N;absurdidad N;esperpento N;ridiculez N;abundancia P+;maremagno P+;magno P+;maremágnum P+;abundante P+;copioso P+;abusar N+;abuso N+;arbitrariedad N+;extralimitación N+;abismo N;abismar N;acceder P;acelerar +;aceleración +;aceleramiento +;acentuar P;aceptar P;admitir P;aceptación P;accesible P;certero P;fiel P;preciso P;precisar P;precisión P;adquirir +;adquisición +;activo +;activar +;adaptar +;pacotilla N;adecuar +;apropiar +;adaptabilidad P+;acomodable P+;exagerar N;adaptable P+;adaptación P+;amoldamiento P+;adaptativo P+;agregar +;añadir +;incorporar +;sumar +;añadidura +;suma +;sumar +;adicional +;suplementario +;ajustar P+;graduable P;orientable P;regulable P;adecuación P;ajuste P;ajustar P;arreglo P;arreglar P;retoque P;retocar P;administrar +;administrador +;admirable P;admiración P+;admirar P;admirador P+;adorable P;adorar P;adornar P+;adorno P;adornar P;adulterar N;adulteración N;adelantar P+;adelanto P+;anticipado P+;anticipar P+;avance P+;aprovechamiento P+;aventajamiento P+;bonificación P+;delantera P+;ventaja P+;ventajoso P+;venturoso P+;adverso N;negativo N;adversidad N;aconsejable P;conveniente P;afabilidad P;agrado P;agradar P;afectar +;cariño P;simpatía P;afectuoso P;cariñoso P;afinidad P+;afirmar P+;constatar P+;afirmación P;afirmativo P;afligir N+;atribular N+;aflicción N-;afligimiento N-;afluencia P;temeroso N-;agravar N+;enconar N+;agravación N+;agregado P;agregar P;agregación P;ágil P+;agilidad P+;agonía N;agradable P;asombroso P+;desconcertante N+-;pasmoso N+-;ambigüedad N;equívoco N;ambiguo N;ambicioso +;ameno P;amabilidad P;amable P;amistoso P;amplio P+;divertir P;entretenimiento P;antiguo -;ángel P;angelical P;coraje N;encolerizar N;enojar N;enojo N;rabia N;airado N;airar N;fastidiar N;enfado N;enfadar N;fastidio N;fastidiar N;molestia N;anómalo N;anomalía N;antipatía N;ansia N-;ansiedad N-;ansioso N-;inquieto N-;inquietar N-;apático N-;apatía N-;apreciar P;apropiado P;oportuno P;visto bueno P;aprobar P+;apto P+;arduo +;peliagudo N;trabajoso N;suavidad P;arrogancia N;altanero N+;altivo N+;arrogante N+;petulante N+;arte P;artificial N;astuto P;atlético P+;sobre todo +;atroz N;atento P;agraciado P;atractivo P;atrayente P;audaz N+;aumentar +;austero P-;auténtica P;autenticidad P;avaricia N;avaro N;medio -;aversión N;terrible N+;malo N;mal N;balance +;equilibrado +;bancarrota N;quiebra N;quebrar N;bárbaro +;barrera N;bestial +;descubrimiento P;lúcido P;bello P;artista P;hermoso P;beldad P;belleza P;primor P;beneficioso P;provechoso P;beneficiario P;beneficiar P;beneficio P;provecho P+;benevolencia P;benévolo P;benevolente P;benigno P;irrenunciable P;espiritual P;bueno

P;mejor +;decepcionante N+;mejorar P;agrio N;amargo N;amargar N;amargado N;acritud N;amargura N;florecer P;atrevido +;atrever +;audacia P+;desgarrador +;boom P;aburrir N+;aburrimiento N;tedio N;molestar N;bravo P+;valiente P+;braveza P+;valor P+;incumplimiento N+;romper N+-;genuino +;muy +;alegre P;alegrar P;lustroso P;brillo P;brillar P;claridad P;brillantez P;fulguración P;resplandor P;brillante P;genial P;reluciente P;obramaestra P+;gran +;boyante P+;grave N-;calamidad N;apacible P;calma P;quietud P;tranquilo P;sosiego P;capaz P+;hábil P+;capacidad +;antojadizo N;caprichoso N;cautivar P-;cuidar P;esmero P;cuidadoso P;dejadez N;acariciar P;caricia P;catástrofe N;provocar +;celebridad P;campeón P+;caos N;caótico N;carisma P+;caritativo P;caridad P;encanto P;encantar P;casto P;puro P;barato N-;tramposo N;alegrar P;alegre P;chic P;maestría P+;moroso N;amoroso P-;claridad P;clásico P;limpio P;limpieza P;claro P;definido P;nítido P;preciso P;sensible P;sencillo +;transparente P;ni -;nitidez P;capaz P+;cercano P;entrañable P;torpe N;basto N;burdo N;tosco N;bastedad N;basteza N;coherente P;frío N;colosal P+;cómodo P;comodidad P+;confort P+;reconfortar P+;encomiable P;alaordendeldía P;común P;corriente P;conmoción N;jaleo N;revuelo N;opulento N;también +;compasión P;afín P;compatible P;compensar P-;recompensar P;competente P+;completo +;complejo N;complejidad N;complicación N;concreto +;confianza P+;seguridad P+;felicitar P;considerable +;estimable +;continuo +;contener +;despreciar N;menospreciar N;rechazo N;rechazar N;menospreciable N;despectivo N;contento P;contentar P;contradicción N;contradictorio N;polémico N;controversia N;ingenuo N-;irreconocible N;engendro N+;delicioso P;petardo N;supuesto -;correcto P;corrección P;corrosivo N;corrupto N;corrupción N;valor P+;valiente P+;cortés P;cortesía P;moda N-;desorbitado N;disparatado N;creación +;creativo P;potingue N;creatividad P;credibilidad P;caricatura N-;creíble P;crítico N;criticar N;grosero N;cruel N;inhumano N;siempre +;crueldad N;abrumar N+;seco N;cínico N;cinismo N;deterioro N+;peligroso N;arriesgado N;oscuro N;engaño N;decencia P;decente P;decoroso P;digno P;decepción N;engañoso N;falaz N;decisivo +;determinante +;simple +-;decorativo P;decrecer -;decrecimiento -;mermar -;roto N;defecto N-;fallo N-;falto N-;defectuoso N-;deficiencia N-;deficiente N-;deficitario N-;déficit N-;delicado P-;delicia P;encanto P;depreciación N;depresión N-;descender -;merecido P+;deseable P;escaso -;previsible N-;despreciable N;suerte +;perjudicial N+;carente N-;difícil N;dificultoso N;espinoso N;grueso N;dificultad N;dilema N;disminuir -;menguar -;minimizar -;rebajar -;directo +;sucio N;desventaja N-;inconveniente N-;decepcionar N;desilusionar N;chasco N;desengaño N;desengañar N;desilusión N;desastre N;desastroso N-;alto +;incomodo N;discreto P;discrepante N;desprecio N;asco N+;penoso N+;deshonesto N-;deshonra N-;rabia N;triste N-;disgustar N-;desagrado N-;desagradar N-;descontento N-;insatisfacción N;peculiar P;distinguido P+;distorsión N;dividir N+-;duda N;dubitativo N-;dudoso N-;indudable P+;ruina N;pavoroso N-;bajada -;bajar -;insulso N-;dinámico P+;facilidad P;estrafalario N;excéntrico N;excentricidad N;efectivo P;eficaz P+;poderoso P+;vigente P+;efectividad P+;eficacia P+;eficiencia P+;eficiente P+;estudiado P;elaborado P;trabajado P;elegancia P;elegante P;elevar +;embellecer P;énfasis +;enfático +;rotundo +;vacío N-;encantar P+;encanto

P;inacabable +;disfrutar P;disfrute P;deleitable P;goce P;brutal +;enorme +;ingente +;entretenimiento P;entusiasmo P;entusiasmar P;equitativo P;error N;prescindible N;imprescindible P+;preciso P+;eterno +;ético P;nunca -;exacto P+;fiel P+;exceder +;sobrepasar +;sobre +;sobresalir P+;excelencia P;impresión +;excelente P+;relevante P;excepción N;exceso N;desmedido +;excesivo N;entusiasmar P;exclusivo P;depurado P;exótico P;ampliar +;expandir +;expansionar +;dilatación +;expansión +;pánfilo N;caro N;costoso N;experto P+;fenomenal P+;ampliación +;extensión +;extenso +;descomunal +;extraordinario P;extremado +;fabuloso P;caro N;devastador N+;joya P+;facilidad +;conveniente P;fiel P;bajada -;bajar -;caer -;caída -;descenso -;célebre P;famoso P;renombrado P;fantástico P;farsa N;fascinación P;alamoda P;fatal N;favorable P;favorito P;predilecto P;festivo P;fiasco N-;digno P;halago P-;zalamería P-;desperfecto N-;huida N-;huido N-;fugaz N;correoso P;flexible P;creciente +;majadero N;necio N-;desatinado N-;insensato N-;formalidad P;bienaventurado P;feliz P;frágil N-;quebradizo N-;achacoso -;campechano P;franco +;defraudación N-;fraudulento N-;desaprovechado N;gratis P+;gratuito P+;libertad P+;libre P+;amigable P;amistoso P;amistad P;espantar N;aterrador N;espantoso N;temible N;frígido N;frívolo N;fructífero P+;completo +;lleno +;pleno +;rotundo +;plenitud P;fundamental +;primordial +;furioso N;delicado N;ganancia P+;galán P+;dadivoso P;generoso P;genialidad P+;genio P+;gigante +;facilidad P;dotado P+;gigantesco +;alegre P;alegrar P;glamour P;deslumbramiento N+;fosco N;lóbrego N;triste N;glorioso P;gloria P;bueno P;conveniente P;digno +;bodrio N+;bello P;gracia P;grácil P;gracioso P;lucido P;grandeza P+;gratitud P;gratuito N;pésimo N+;redundante N;esperpéntico N;sorprendente P;crecimiento +;culpable N-;profesionalidad P;honestidad P;logrado P-;confuso N-;guapo P;felicidad P;dichoso P;feliz P;inverosímil N;duro N+-;fuerte N+-;tieso N+-;dañino N+;lesivo N+;armonioso P;armonía P;odiar N;aborrecedor N;aborrecimiento N;odio N;peligro N;calinoso N;saludable P;sano P+;desalmado N;descorazonador N;héroe P+;heroico P+;alto +;elevado +;estorbo N;hueco N-;sagrado P;honesto P;honrado P;decoro P;honor P;honorable P;honroso P;horrendo N;horrible N;horror N;monumental +;humanitario P;humilde P-;humildad P;jocoso P;colgado N-;dolido N+;herido N+;hipócrita N;histerismo N;ideal P;idiota N;tonto N;ociosidad N;ignorancia N-;ilegal N;indebido N;ilegalidad N;analfabeto N-;iletrado N-;antilógico N-;ilógico N-;imaginación P;imaginativo P;imitación N;inmaculado P;inmaduro N-;inmenso +;inmoral N;inmoralidad N;inmovible N+;inmóvil N+;imparcial P;imparcialidad P;impaciente N;impedimento N;imperfecto N;impersonal N;impetuoso N;importancia P;trascendencia P;importante P;relevante P;trascendente P;impactante P+;impresionante P+;impropio N;mejora P+;mejoría P+;inaccesible N;inexactitud N;inadecuado N-;adecuado P-;selecto P+;horroroso N+;incapaz N-;constante N+;incesante N+;incompatible N;incapacidad N;incompetente N;incompleto -;inconsistencia N;inconveniente N;incorrecto N-;incrementar +;incremento +;increíble N+-;incurable N;irresoluble -;indefinido N;independiente P+;inefable P;indicativo P;indiferencia N;indiferente N;indignación N;indirecto -;exigible P+;indispensable P+;indisputable +;individualidad P;indomable P+;indulgencia P;ineficaz N;ineficacia N;ineficiencia N;desigualdad N;inevitable

+;inexacto N;inexplicable N;infalible P+;inferior N-;inferioridad N-;infinito +;hinchado N;inflación N;influencia +;influyente +;ingenioso P;inherente +;inhibición N+;iniciativo +;nocivo N-;inocencia P;cándido P;candoroso P;inculpable P;inocente P;innumerable +;curioso P;inquisitivo P;inseguridad N-;insensible N;inseparable P;insignificante N-;nimio N-;insistente P+;chulería N;chulo N;insolente N;instintivo P;insuficiencia N-;insuficiente N-;insulto N;intacto +;intelectual P+;inteligencia P+;inteligente P+;inteligible P;intenso +;intensidad +;intensivo +;interés P;interesado P;interesar P;íntimo P;intolerable N-;nulo N;invariable +;invencible P+;involuntario N;invulnerable P+;irónico N;ironía N;irracional N;irrefutable +;irregular N;irregularidad N;irresistible P;irresponsable N-;enfadadizo N;gracia P;alegría P;gozo P;júbilo P;gozoso P;jubiloso P;justo P;preciso P;amable P;atento P;bondadoso P;bondad P;detalle P;falto N-;amplio +;grande +;apabullante N+-;último +;enjuto -;legal P;legítimo P;responsable N;liberal P;claridad P-;leve P-;ligero +-;liviano P-;tampoco -;poco -;animado P;marchoso P+;vital P+;lógico P;alto +;hermoso P;bajar -;fiel P;leal P;lealtad P;moderno P;suerte P+;ventura P+;ganancial P;lucrativo P;luminoso P;lujoso P;lujo P;magnífico P+;soberbio P+;principal +;majestuoso P;majestuosidad P;generalidad +;mayoría +;mayoritario +;inadaptación N-;malicioso N;maligno N;manejable P+;varonil P+;marginal N;marital P;amar +;más +;comedido +;mediocre N-;memorable P;peligro N+;clemente P;misericordioso P;clemencia P;misericordia P;piedad P;excepcional P+;mero -;mérito P+;lío N;zafarrancho N;metódico +;meticuloso P;minucioso P;caudaloso P+;menor -;milagro P;milagroso P;cutre N;miserable N-;mísero N-;miseria N-;desdicha N-;desgracia N-;desventura N-;percance N;equivocación N;error N;fallo N;equivocado N;suspicacia N;incomprendido N;ridiculización N;moderado P;moderación P;modernidad P;humilde P-;atribulado N+;modesto P-;púdico P-;momentáneo -;monstruo N+;monstruoso N+;monumento +;monumental P;moral P;moralidad P;motivación P;emocionar +;barroso N;cenagoso N;legamoso N;limoso N;lodoso N;musculoso +;angosto -;estrecho -;natural P;curioso P;necesario P;preciso +;fiel N;descuido N;negligencia N;negligente N;nervioso N-;nerviosismo N-;respeto N-;neurótico N;nuevo +;amable P;pesadilla N;caballeroso P+;noble P;estrépito N;violento +;despropósito N;majadería N;necedad N;pamplina N;tenue -;sutil -;notable P+;notoriedad P;novedad P;novato N-;nutriente P;oscuro N;obsoleto N-;obstáculo N+;reacio N;terco N;entorpecimiento N;obstrucción N;taponamiento N;curioso N;peculiar N;sabio P+;raro N;insultante N+;ofensivo N+;oficial +;anciano -;antiguo -;viejo -;nefasto N+;ominoso N+;solo -;insufrible N+;tan +;tansolo -;único +-;oportuno P;oportunidad P;agobio N;óptimo P;optimismo P;originalidad P;destacado +;relevante P+;nada -;abrumador N;agobiador N;aplastante N;avasallador N;deslumbrante N+-;propio +;pena N;doloroso N;pálido -;paraíso P;parásito N;gracia P+;rocambolesco N-;pasión +;pasivo -;patético N+;particular N;peculiar N;mierda N+;perfección P+;apasionante P+;perfecto P+;permanente +;perplejo N-;perplejidad N;persistente +;pertinente P;avieso N;piadoso P;lastimoso N-;peste N+;plaga N+;lúdico P;agradable P;ameno P;gustar P;agradar P;complacer P+;gustado P+;venenoso N;positivo P;potencial +;potencialidad +;virtud P;poderoso +;potente +;practicable P;elogio P;precario N-;precioso P+;preciso P;puntual P;precisión P;prematuro -

;principio P;prestigioso P+;cursi N;bonito P;cuco P;lindo P;intimidad P;privacidad P;privilegio +;pro P;dificultad N;problema N;prodigioso P;portento P;prodigio P;beneficio P+;logro P+;lucro P+;rentable P+;progreso P+;progresivo P;próspero P;orgulloso P+;ufano P+;exquisitez P+;puro P;pureza P;calidad P;calmoso +;quieto +;silencioso +;menos -;coherencia +-;aburrido N;real P;antológico P;verdadero P;asequible P;moderado P;plausible P;razonable P;revoltoso N+;insumiso N;rebaje -;reducir -;restar -;descenso -;reducción -;redundancia N;regular +-;estrella +;comercial N-;fiabilidad P+;seriedad P+;solvencia P+;confiable P+;cumplidor P+;serio P+;solvente P+;destacable P+;notable P+;remoto -;respetable P;respetuoso P;responsable P;revolución P;revolucionario +;rico P+;ridículo N-;rudo N;áspero N;basto N;despiadado N;sagrado P;lastimero N;triste N;tristeza N;escrupuloso N;sedentario N;sensacional P+;prudente P;sensato P;sensible P;quisquilloso P;sensibilidad P;sentimental P;formal P;serio P;seriedad P;severo N+;severidad N+;sombrío N;umbrío N;vergüenza N;inconfesable N;vergonzoso N;demasiado +;mejorable N-;pastiche N+;corto N-;españolada N;trascendencia P+;significante P+;significativo P+;sencillez P;simplicidad P;cero -;sincero P;sinceridad P;pecaminoso N;arte P;destreza P;habilidad P;pericia P;habilidoso P;magistral P;calumnioso N;majo P+;sonrisa P;sollozo N;sociable P;consistente +;sólido +;solidez P+;desobra +;especial P;aparatoso N;espectacular P+;velocidad P;espléndido P;esplendor P;deslucir N;estabilidad P+;estable P+;estático N;aún +;todavía +;poderoso +;profundo +;recio +;bobo N;estúpido N;estupidez N;torpeza N;substancial +;sustancial +;sustancioso +;sutil P-;acierto P;éxito P;exitoso P;superficial N;superfluo N;superior P+;superioridad P+;superlativo P;solidario P;dulce P;mimoso P;talento P+;talentoso P+;alto +;tenaz P+;tenacidad P+;tesón P+;ternura P;consistente +;espeso +;grueso +;espinoso N;desatento N;ahorrativo P;diminuto -;flexible P;tolerante P;apoteósico P+;tremendo +;dificultad N;verdadero P;fiable P+;emocionante P+;realidad P;verdad P;veraz P;antiestético N;estético P;humanístico P;sensual P;feo N;forzoso N;insoportable N;incierto -;inseguro -;incertidumbre -;inseguridad -;incómodo N;molesto N;indeseable N;desfavorable N;inacabado -;inconcluso -;inolvidable P;desafortunado N-;desdichado N-;funesto N-;insalubre N-;malsano N-;injusto N;inconfundible +;antinatural N;innecesario N;desapercibido N;antipático N;desapacible N;impopular N;incuestionable +;indiscutible +;insatisfactorio N;desaprensivo N-;indecible N;inestable N-;útil P;inútil N;vago N-;valeroso P;valiente +;válido P;validez P;valioso P;versátil P;viable P;nulo N;vulnerable -;cálido P;cansino N;cordial P;derroche N-;desperdicio N-;despilfarro N-;bien P;maravilla P;maravillar P;sorprender P;estupendo P;formidable P;maravilloso P;célebre P+;famoso P+;renombrado P+;escalofriante N-;cutre N+;casposo N;enfermizo N;peor N+;pesado N;recargado N;

REFERENCIAS

- Andrea Esuli y Fabrizio Sebastiani, "*Determining Term Subjectivity and Term Orientation for Opinion Mining*," 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.

- Bo Pang y Lillian Lee, "*Opinion mining and sentiment analysis*," Foundations and Trends in Information Retrieval, Jinan(China), 18-21 Agosto de 2008, Vol.2, No.1-2, pp.1-135.

- Casey Whitelaw, Navendu Garg y Sholmo Argamon, "*Using Appraisal Taxonomies for Sentiment Analysis,*" MCLC-05, the 2nd Midwest Computational Linguistic Colloquium, 2005.

- Bo Pang y Lillian Lee, "*Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales*," *ACL 2005*, pp. 115-124.

- MariSokolova y Guy Lapalme, "*Verbs Speak Loud: Verb Categories in Learning Polarity and Strength of Opinions,*" Canadian Conference on Al 2008, pp. 320-331.

- Philip Beineke y Trevor Hastie, "*The Sentimental Factor: Improving Review Classification via Human-Provided Information*," 42nd Annual Meeting on Association for Computational Linguistics, Bacelona (España), 2004.

- Pimwadee Chaovalit y Lina Zhou, "*Movie Review Mining: a Comparison between Supervised an Unsupervised Classification Approachs*," 38th Hawaii International Conference on System Sciences, 2005.

- Janyce Wiebe y Rada Mihalcea, "*Word Sense and Subjectivity*," 44th Annual Meeting of the Association for Computational Linguistics, Julio de 2006.

- Nathanael Chambers y Joel Tetreault, "*Approaches for Automatically Tagging Affect,*" Computing Attitude and Affect in Text: Theory and Applications, pp. 143-158.

- Isaac Martín de Diego y Ángel Serrano, "*Técnicas de reconocimiento automático de emociones,*" en Revista Electrónica Teoría de la Educación, Diciembre de 2006, Vol.7, No.4.

- Bo Pang y Lillian Lee, "*Thumbs up? Sentiment Classification using Machine Learning Techniques*," Conference on Empirical Methods in Natural Language Processing (EMNLP), Junio de 2002, pp. 79-86.

- Peter D. Turney, "*Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,*" 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, Julio de 2002, pp. 417-424.

- David García y Francesc Alías, "*Identificación de emociones a partir de texto usando desambiguación semántica,*" Procesamiento del Lenguaje Natural, Revista nº 40, Marzo de 2008, pp. 75-82.

- Alekh Agarwal y Pushpak Bhattacharyya, "*Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of document to be classified,"* International Conference on Natural Language Processing (ICON), 2005.

- Minqing Hu y Bing Liu, "*Mining Opinion Features in Customer Reviews,*" 19th National Conference on Artificial Intelligence (AAAI-2004), San Jose (USA), Julio de 2004.

- Theresa Wilson, Janyce Wiebe y Rebecca Hwa, "*Just how mad are you? Finding strong and weak opinion clauses*," 19th National Conference on Artificial Intelligence (AAAI-04), San Jose (California), Junio de 2004.

- Ellen Riloff, Janyce Wiebe y William Phillips, "*Exploiting Subjectivity Classification to Improve Information Extraction,*" 20th National Conference on Artificial Intelligence (AAAI-05), Junio de 2005.

- Carlos G. Figuerola, Jose L. Alonso Berrocal y Angel F. Zazo Rodriguez, "*Algunas Técnicas de Clasificación Automática de Documentos*," Cuadernos de documentación multimedia, ISSN 1575-9733, Nº. 15, 2004 , pp. 1-2.

- Kagan Tumer y Joydeep Ghosh, "*Order Statistics Combiners for Neural Classifiers*," World Congress on Neural Networks, 1995.

- Xue Bai y Rema Padman, "*On learning Parsimonious Models for Extracting Consumer Opinions,"* 38th Hawaii International Conference on System Sciences (HICSS'05), Vol.03, Enero de 2005.

- Francisco José Cortijo Bon, "*Técnicas supervisadas II: Aproximación no paramétrica,"* Universidad de Granada, Octubre de 2001.

- Hugo Liu, Henry Lieberman, y Ted Selker, 2003, "*A model of textual affect sensing using real-world knowledge,"* 8th International conference on Intelligent user interfaces, Miami(Florida, USA), 2003, pp. 125-132.

- Enrique Cabello Pardos, "*Técnicas de reconocimiento facial mediante redes neuronales*," Tesis doctoral, Facutad de Informática (UPM), Abril de 2004.

- Hrvoje Bacan, Igor S. Pandzic y Darko Gulija, "*Automated News Item Categorization,"* 2005.

- Virginia Francisco y Pablo Gervás, "*Análisis de dependencias para la marcación de cuentos con emociones*," Procesamiento del lenguaje natural, ISSN 1135-5948, Nº. 37, 2006 , pp. 137-144.

- Eui-Hong (Sam) Han, George Karypis y Vipin Kumar, "*Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification*," 1999.

- Fabrizio Sebastiani, "*Machine Learning in Automated Text Categorization,*" Consiglio Nazionale delle Ricerche, Italia, Junio 2002.

- Vladimir N. Vapnik, "*The Nature of Statistical Learning Theory*," 1995.

- Soo-Min Kim y Eduard Hovy , "*Identifying and Analyzing Judgment opinions*," Human Language Technology Conference, New York, 2006, pp. 200-207.

- Alekh Agarwal y Pushpak Bhattacharyya, "*Sentiment Analysis: A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified,*" International Conference on Natural Language Processing(ICON), IIT Kanpur, India, Diciembre de 2005

- Keke Cai, Scott Spangler, Ying Chen, Li Zhang, "*Leveraging Sentiment Analysis for Topic Detection*," International Conference on Web Intelligence and Intelligent Agent Technology, 2003, vol. 1, pp.265-271.